Universidade do Minho
Escola de Engenharia

Tao Yu

AI-assisted Teamwork Grouping for Elementary Students' STEAM Project-Based Learning

Fevereiro de 2026

AI-assisted Teamwork Grouping for Elementary Students' STEAM Project-Based Learning
Tao Yu
UMinho | 2026

Universidade do Minho
Escola de Engenharia

Tao Yu

AI-assisted Teamwork Grouping for Elementary Students' STEAM Project-Based Learning

Tese de Doutoramento
Programa Doutoral em Engenharia Eletrónica e Computadores

Trabalho efetuado sob a orientação de:
Professor Doutor Sandro Emanuel Salgado Pinto
Professor Doutor Tiago Manuel Ribeiro Gomes
Professor Doutor Hao Xu

Fevereiro de 2026

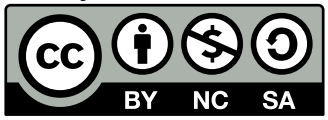
COPYRIGHT AND TERMS OF USE OF THIS WORK BY A THIRD PARTY

This is academic work that can be used by third parties as long as internationally accepted rules and good practices regarding copyright and related rights are respected.

Accordingly, this work may be used under the license provided below.

If the user needs permission to make use of the work under conditions not provided for in the indicated licensing, they should contact the author through the RepositoriUM of Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International
CC BY-NC-SA 4.0

`https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en`

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.0.2) .

Acknowledgements

As this work reaches completion, I look back on my years of study and research in Portugal with deep gratitude, aware that this journey has never been the work of one person alone.

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. AT, Prof. Hao XU, Prof. Sandro, and Prof. Tiago. Their rigorous guidance, insightful feedback, and unwavering support have been invaluable throughout my research. Beyond academic mentorship, they have also shaped my perspective and resilience, encouraging me to pursue my aspirations with perseverance.

I am profoundly grateful to my parents for their unconditional support and understanding. I am also deeply aware of my long absence and the limited time I have been able to spend with them over the years. I hope that the completion of this study can, in some small way, repay their devotion and sacrifices.

I would also like to thank my Portuguese friends, Chinese friends, and everyone I met here. Your companionship in the laboratory, as well as your support in daily life, has enriched my experience in Portugal and made this journey far more meaningful.

My special thanks go to my dear Irina. Thank you for your care, understanding, and the genuine engagement you have offered me, which has brought warmth and a sense of home during challenging times. Love you.

Finally, I would like to thank myself for staying the course, for not giving up, and for continuing to move forward until reaching this point.

May everyone hold on to their dreams and work earnestly toward them. I wish all those who have accompanied and supported me the very best.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

_____, _____

(Place) (Date)

_____

(Tao Yu)

"*Journey of a thousand miles begins with single step.*"

— Lao Tzu, Tao Te Ching
(Author of the Tao Te Ching, one of the foundational texts of Taoism)

Resumo

Um mecanismo de agrupamento de equipas assistido por IA para a aprendizagem STEAM baseada em projetos no ensino básico

O trabalho colaborativo em grupo é central na aprendizagem STEAM no ensino básico em PBL, influenciando participação, apoio entre pares e equidade. Contudo, o agrupamento em sala de aula é difícil de tornar reprodutível e explicável. Os professores decidem sob pressão e rapidamente com evidência incompleta e heterogénea, e com dados multimodais (áudio–vídeo, transcrições, questionários e rubricas) frequentemente ruidosos, em falta e desalinhados. Um problema crítico é que muitas abordagens assumem a fiabilidade uniforme da evidência, pelo que pequenos erros de medição podem distorcer desproporcionalmente os resultados, reduzindo estabilidade e explicabilidade.

Este doutoramento propõe um enquadramento de agrupamento inteligente sensível ao alinhamento que organiza observações em unidades de evidência rastreáveis, estima a fiabilidade via alinhamento intermodal e aprende representações de alunos com uma rede neuronal de grafos. O agrupamento é formulado como otimização multiobjetivo com restrições, garantindo exequibilidade na sala de aula e equilibrando complementaridade, potencial de colaboração e equidade intergrupos, com justificações baseadas em evidências e ajustadas pelo professor. As contribuições incluem: (1) modelação de fiabilidade orientada para a decisão a partir do alinhamento multimodal; (2) modelação interpretável do aluno baseada em multimodalidade e grafos; (3) uma formulação de agrupamento multiobjetivo operacional com compromissos ajustáveis; (4) um desenho auditável orientado para o professor que acompanha as decisões até às unidades de evidência. Sobre dados reais recolhidos nas salas de aula, o método proposto obteve o melhor desempenho nas rubricas do professor face ao agrupamento heurístico professor e ao agrupamento aleatório, mantendo uma baixa disparidade entre os grupos e melhor cumprimento dos objetivos auditados em termos de equidade e potencial colaboração.

Palavras-chave: Agrupamento Inteligente, Alinhamento Intermodal, Formação de Grupos, Modelos de Linguagem de Grande Escala Multimodais, Otimização Multiobjetivo

Abstract

AI-assisted Teamwork Grouping for Elementary Students' STEAM Project-Based Learning

Collaborative group work is a core mechanism in primary-school STEAM project-based learning (PBL), shaping participation, peer support, and equitable learning opportunities. Yet classroom grouping remains hard to make reproducible and explainable because teachers must decide quickly with incomplete, heterogeneous, and dynamically changing evidence, while multimodal classroom data (audio–video, transcripts, questionnaires, and rubrics) are often noisy, missing, and misaligned. The key gap is that existing approaches usually treat all evidence as equally reliable. As a result, when classroom data are noisy or partially missing, minor measurement errors can disproportionately affect grouping outcomes, reducing their stability and explainability.

This Ph.D. work proposes an alignment-aware intelligent grouping framework that organizes heterogeneous observations into traceable evidence units, estimates evidence reliability from cross-modal alignment, and learns interaction-aware learner representations using a graph neural network. It then formulates grouping as a constrained multiobjective optimization problem that satisfies classroom feasibility while balancing complementarity, collaboration potential, and between-group fairness, which provides evidence-linked rationales for teacher-in-the-loop adjustment. Contributions include: (1) decision-oriented reliability modeling from cross-modal alignment; (2) an integrated multimodal model and graph-based approach for interpretable learner modeling; (3) a deployable constrained multiobjective grouping formulation with tunable trade-offs; (4) an auditable teacher-facing design that traces decisions back to evidence units. Results on the real classroom dataset show that the proposed intelligent grouping achieved the highest teacher-rubric performance compared with teacher heuristic grouping and random grouping, while maintaining low between-group disparity and stronger audited objective attainment in fairness and collaboration potential.

Keywords: Cross-Modal alignment, Group Formation, Intelligent Grouping, Multimodal Large Language Models (MLLM), Multi-Objective Optimization

Contents

List of Figures **xi**

List of Tables **xii**

1 Introduction **1**

1.1 Motivation . . . 1
1.2 Problem Statement and Challenges . . . 3
1.3 Scope . . . 5
1.4 Research Questions . . . 6
1.5 Thesis Outline . . . 7
1.6 List of Publications . . . 8

2 Background **9**

2.1 STEAM PBL in Primary School . . . 9
2.1.1 Teamwork of Elementary Students in PBL . . . 11
2.1.2 Educational Digitalization and AI Trends . . . 14
2.2 MLLM and Graph Neural Networks (GNN) . . . 16
2.2.1 Multimodal Data and MLLM in Education . . . 16
2.2.2 Graph-based Model Applications in Education . . . 17
2.2.3 MLLM and GNN for Grouping . . . 18
2.3 Collaborative Grouping and Multiobjective Optimization . . . 19
2.3.1 From Manual to Intelligent Grouping . . . 20
2.3.2 Algorithms and Standards for Grouping . . . 20
2.3.3 Multiobjective Grouping and Constraints . . . 21

3 State-of-the-Art **23**

3.1 Group Formation . . . 23
3.2 MLLMs in Education . . . 25
3.3 Graph-Based Modeling . . . 26
3.4 Multiobjective Metaheuristics . . . 27
3.5 Research Gap . . . 28

4 Graph-enhanced Mutual Learning for Text Classification **30**

4.1 Motivation . . . 30
4.2 GEML Framework . . . 32
4.2.1 Problem Setting . . . 32
4.2.2 Mutual Learning via Probability Distributions . . . 33
4.2.3 Asymmetrical Learning . . . 34
4.2.4 Uncertainty-Weighted Loss Composition . . . 34
4.2.5 Optimization with Document–Node Bijection . . . 35
4.3 Experimental Evaluation . . . 35
4.3.1 Datasets . . . 36
4.3.2 Baselines and Implementation Details . . . 36
4.3.3 Overall Results . . . 38
4.4 Discussion and Connection to the Thesis . . . 39

5 CF-STGAT: Clustered Federated Spatio-Temporal Graph Attention Networks 41
5.1 Motivation . . . 41
5.2 CF-STGAT Framework . . . 43
5.2.1 Preliminaries . . . 44
5.2.2 Dynamic clustering and cluster-wise aggregation . . . 44
5.2.3 Similarity-weighted inter-cluster fusion . . . 45
5.3 Experimental Evaluation . . . 46
5.3.1 Datasets . . . 46
5.3.2 Implementation details and Overall Results . . . 46
5.3.3 Clustering stability . . . 47
5.3.4 Mechanistic Interpretation . . . 48
5.4 Discussion and Connection to the Thesis . . . 49

6 Experiments for STEAM PBL Classroom Data Collection 51
6.1 Experimental Preparation . . . 51
6.1.1 Roles and Responsibilities . . . 52
6.1.2 Four STEAM PBL Projects . . . 53
6.1.3 Control and Treatments . . . 55
6.1.4 Data Infrastructure and Questionnaires . . . 56
6.2 Experimental Data . . . 60
6.2.1 Raw Data Archiving . . . 61
6.2.2 Metadata Indexing . . . 64
6.2.3 Data Annotation . . . 65
6.2.4 Standardized Packaging . . . 65

7 Methodology for Intelligent Group Formation in PBL **67**
7.1 Problem Formulation . . . 67
7.2 Data Sources and Evidence Units . . . 69
7.2.1 Classroom Audio-Video and Transcript Evidence . . . 69
7.2.2 Questionnaire-Based Learner Trait Evidence . . . 70
7.2.3 Teacher Evaluation Rubric Evidence . . . 71
7.3 Cross-Modal Alignment Statistics . . . 71
7.4 Feature Engineering . . . 72
7.4.1 Learner-Level Behavioral and Trait Features . . . 72
7.4.2 Interaction-Oriented Features, Normalization, and Structured Representation 74
7.5 Interaction Graph Construction . . . 76
7.5.1 Edge Extraction from Reply Links . . . 76
7.5.2 Multi-Evidence Edge Weight Fusion . . . 77
7.6 Graph Neural Modeling for Structure-Aware Grouping . . . 77
7.6.1 GNN Objective and Outputs . . . 77
7.6.2 Attention-Based Aggregation and Structural Descriptors . . . 78
7.7 Constrained Multi-Objective Group Optimization . . . 79
7.7.1 Decision Variables and Feasible Set . . . 79
7.7.2 Objective Functions . . . 80
7.7.3 Multi-Objective Aggregation and Solver . . . 81
7.7.4 Auditability and Teacher Intervention . . . 81

8 Evaluation **82**
8.1 Experiments . . . 82
8.1.1 Datasets . . . 82
8.1.2 Baselines and Compared Methods . . . 83
8.2 Results and Analysis . . . 85
8.2.1 Overall Results . . . 85
8.2.2 Framework-level Performance . . . 86
8.3 Summary . . . 91

9 Conclusions **93**
9.1 Responses to Research Questions . . . 93
9.2 Contributions . . . 95
9.3 Limitations . . . 96
9.4 Future Work . . . 97

Bibliography **99**

List of Figures

1 Multimodal classroom context and evidence sources in STEAM PBL. 2
2 Intelligent group formation scope in STEAM PBL. 5

3 Model framework of GEML. 32
4 A bijection example. 35
5 Mutual learning results. 38

6 Overview of Fed-STGAT. 42
7 CF-STGAT pipeline. 44
8 Row-normalized average transition matrices across consecutive checkpoints. 48
9 Per-client training accuracy vs. communication rounds on NTU RGB+D 60 (X-Sub). . . . 49
10 Test loss vs. communication rounds. 49

11 Overview of the proposed alignment-aware intelligent group formation framework. . . . 68

12 Alignment confidence derived from cross-modal attention and alignment statistics. . . . 87
13 Group students interaction graph. 88
14 Pre-post radar profiles of students in proposed intelligent grouping. 89
15 Group Scheme. 90

List of Tables

1 Statistics of datasets. 36
2 Accuracy results on different datasets (%). 37
3 Macro-average F1 results on different datasets (%). 37

4 Statistics of datasets. 46
5 Overall comparison (Top-1 accuracy, %). 47
6 Round-to-round clustering stability on NTU 60 (X-Sub) 47

7 Roles and responsibilities. 52
8 Member roles and responsibilities in a group. 53
9 Key index. 57
10 Audio transcript Utterance-Level schema. 58
11 Interaction edge schema. 58
12 Pre-Study Questionnaire (Part I) 60
13 Pre-Study Questionnaire (Part II) 61
14 Teacher rating rubric. 64
15 Standardized dataset schema. 66

16 Summary statistics of the classroom STEAM PBL evidence-chain dataset. 84
17 Main comparison of overall effectiveness and equity. 85
18 Dimension-wise normalized teacher rubric scores. 86
19 Objective-attainment auditing of the final groupings. 90

1 Introduction

In recent years, basic education worldwide has been undergoing a rapid digital transformation. With the development of information technology, the traditional "blackboard + textbook" teaching model is increasingly complemented by data-informed and intelligence-supported learning paradigms that enable more adaptive instructional decisions. Many countries have introduced policies to promote modern, student-centered, data-driven education systems, within which artificial intelligence (AI) has become an important force reshaping educational practice. With the advancement of Multimodal Large Language Model (MLLM), it has become feasible to conduct large-scale analyses of heterogeneous evidence from authentic classroom settings via cross-modal alignment and consistency modeling, and to quantify and propagate evidence reliability under complex conditions. This capability is particularly critical for learning analytics and instructional interventions, because related decisions must be grounded in observable evidence while also meeting requirements for transparency and auditability. Meanwhile, interdisciplinary integration has been permeating basic education curricula. As a representative model, STEAM education integrates Science, Technology, Engineering, Arts, and Mathematics, and emphasizes project-based and collaborative learning. In STEAM project-based learning (PBL), teamwork is not only an organizational arrangement for task completion, but a primary mechanism through which students develop communication, coordination, and collective problem-solving capacities. However, forming student groups scientifically and efficiently remains a persistent challenge in classroom practice. Against the backdrop of advancing AI, this doctoral project explores an alignment-aware approach to intelligent group formation, which integrates educational objectives with evidence-driven modeling and decision optimization.

This chapter begins by presenting the motivation for exploring AI-enhanced student grouping in STEAM PBL education in Section 1.1. Section 1.2 reviews the current challenges in group formation, including the limitations of traditional manual grouping and the need for MLLM. Section 1.3 delineates the scope of this doctoral project, focusing on cross-multimodal data modeling and intelligent group optimization. Section 1.4 outlines the core research questions and the methodological framework adopted to address them. Finally, Section 1.5 provides an overview of the thesis structure, and Section 1.6 summarizes the main academic contributions of this doctoral work.

1.1 Motivation

In the early twenty-first century, international organizations and national education authorities advanced the concept of future-oriented education, emphasizing that basic education should not merely transmit

foundational knowledge but also cultivate students' problem-solving competence, interdisciplinary thinking, and collaborative skills [1]. Particularly at the elementary level, students undergo a developmental transition from dependence toward increasing autonomy, a pivotal period for the formation of self-concept, social identity, and collaborative awareness. STEAM PBL aligns naturally with these developmental characteristics. Primary school students typically demonstrate strong curiosity, a drive to explore, and a desire to express themselves. PBL emphasizes learning through hands-on experiences and the application of knowledge in authentic contexts, engaging students in experimental inquiry, creative design, and iterative refinement.

However, compared with traditional teaching, PBL depends more heavily on effective collaboration within student groups. Whether students are assigned to suitable groups directly affects their engagement, willingness to communicate, task ownership, knowledge uptake, and sense of accomplishment [2]. An effective group is expected to balance complementary capabilities while maintaining a workable level of interpersonal compatibility and coordination. In practice, group formation is still dominated by teacher intuition, often based on perceived academic performance, extroversion, or existing classroom social circles, with limited data support and limited access to latent collaborative relationships reflected in situated interaction. Such experience-based grouping practices can easily lead to ability homogeneity, the concentration of disadvantaged students, within-group communication conflicts, and the neglect of non-academic strengths. These problems motivate the need for a principled grouping mechanism that is fair, effective, and interpretable, while remaining feasible for classroom deployment.

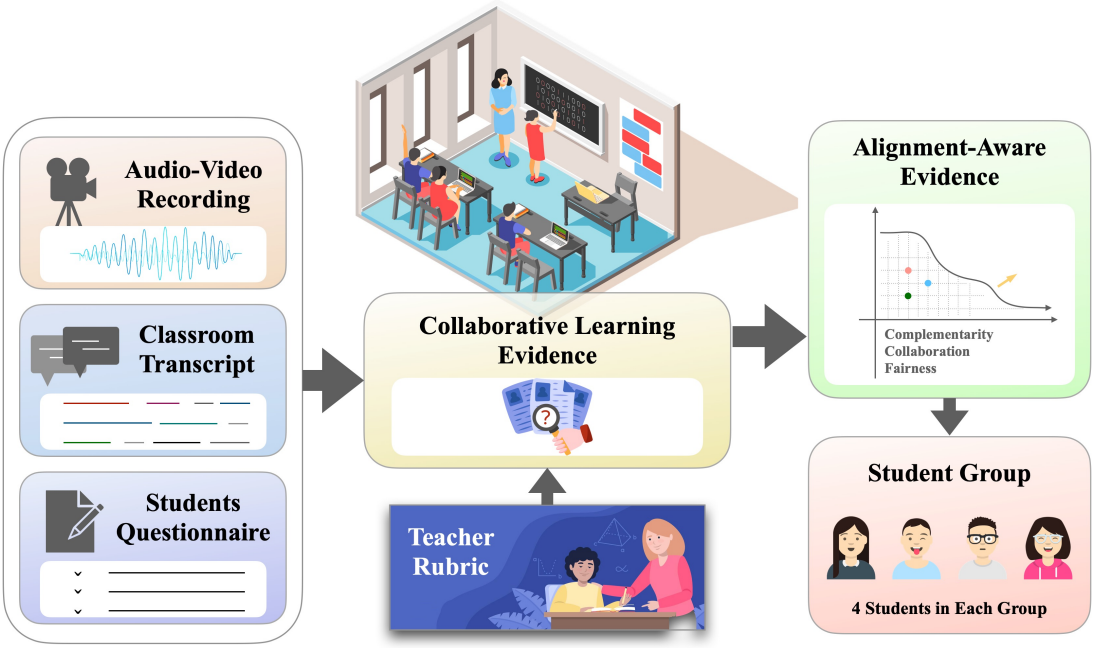


Figure 1: Multimodal classroom context and evidence sources in STEAM PBL.

This research is situated in a Chinese primary school STEAM PBL classroom and leverages heterogeneous evidence sources, including classroom audio-video recordings with aligned transcripts, learner questionnaires, and teacher evaluation rubrics. As illustrated in Figure 1 above, these sources provide complementary perspectives on PBL: recordings capture situated interaction processes, questionnaires reflect relatively stable learner traits and dispositions, and rubrics provide pedagogically grounded outcome assessments. A central methodological challenge is that such multimodal classroom evidence is inherently uncertain; transcripts may be noisy, speech segments may overlap, and the semantic correspondence between a transcript segment and its aligned audio-video window can be imperfect. Accordingly, the intelligent grouping must go beyond simply aggregating signals. It requires transforming heterogeneous observations into reliable indicators, modeling them relationally to represent collaboration as an interaction-dependent process, and explicitly quantifying and propagating evidence reliability. Thus, grouping decisions do not overreact to noisy or weakly supported signals. This study is positioned at the intersection of these educational challenges and advances in AI technologies. It seeks to develop a collaborative grouping framework that integrates alignment-aware multimodal data, graph-structured modeling, and education-oriented objectives. The aim is to provide a high-quality, interpretable, and friendly intervention with intelligent group formation for PBL activities in primary school STEAM PBL settings.

1.2 Problem Statement and Challenges

Although intelligent grouping holds strong theoretical potential, practical deployment in primary school STEAM classrooms faces several structural challenges. These challenges stem not only from algorithmic complexity but also from the mismatch between classroom evidence conditions and the requirements of reliable decision-making. This thesis focuses on three tightly coupled challenges.

First, multimodal evidence exhibits substantial uncertainty and cross-modal misalignment. In real classroom environments, audio–video data are frequently affected by background noise, camera viewpoints, and students' physical movement, leading to distorted speech signals, overlapping utterances, and ambiguous speaker attribution. Meanwhile, transcript data are not consistently accurate, particularly in primary school settings where children's speech characteristics, variable speaking rates, and informal expressions pose additional challenges for automatic transcription systems. More critically, even when temporal alignment between transcripts and audio–video streams is established at the timestamp level, the semantic correspondence between textual content and its associated multimodal context is not guaranteed. Consequently, cross-modal misalignment in classroom data is not merely a matter of temporal synchronization error, but reflects deeper variation in semantic support and evidence reliability. Such variation differs across students, groups, and project phases, evolving dynamically over time. If this heterogeneity in evidence reliability is not explicitly modeled, downstream feature aggregation and relational modeling processes will treat low-quality and high-quality evidence as equally informative, thereby amplifying noise through statistical summaries and graph structures. This amplification can lead to unstable grouping decisions, sensitivity to incidental errors, and reduced reproducibility. Therefore, intelligent

grouping methods must incorporate mechanisms that quantify cross-modal consistency and reliability and propagate these signals throughout feature construction, relational modeling, and decision optimization [3, 4].

Second, collaboration is inherently relational and cannot be adequately characterized by only individual feature similarity. The quality of collaboration in classroom teamwork depends not only on what abilities students possess but on how those abilities are enacted, coordinated, and complemented through interaction. Effective collaboration is typically manifested through patterns such as functional role differentiation, sustained responsiveness and feedback, coordination of task rhythms, and negotiation of perspectives. These mechanisms are shaped by interaction structure and peer influence rather than by static individual traits. For example, two students with similar cognitive or personality profiles may struggle to collaborate productively due to redundant roles or mismatched interaction styles. In contrast, students with distinct strengths may form a highly effective partnership through complementary behaviors. More fundamentally, collaborative potential is often revealed through relational properties such as interaction directionality, frequency, reciprocity, centrality, and brokerage, which cannot be captured by vector similarity alone. These structural properties describe how students influence one another and how coordination opportunities emerge within the group. Accordingly, intelligent grouping requires structure-aware relational modeling that explicitly represents peer interaction topology, rather than treating collaboration as an attribute that can be inferred from isolated learner profiles [5–8].

Third, grouping decisions must satisfy practical classroom feasibility constraints while balancing multiple pedagogical objectives and maintaining interpretability for teacher intervention. Unlike generic clustering or recommendation tasks, classroom group formation is governed by explicit institutional and organizational constraints: group sizes are typically fixed, grouping must be completed within a limited time, and decisions must align with classroom management considerations. At the same time, teachers often hold implicit educational goals for group composition, such as avoiding the concentration of vulnerable students, maintaining approximate balance across groups, managing gender ratios, mitigating interpersonal conflicts, and ensuring that each group possesses a minimal capability to complete the assigned task. In STEAM PBL, grouping must further balance short-term task feasibility with longer-term developmental considerations. For instance, maximizing immediate performance should not come at the cost of suppressing student participation or learning opportunities, while prioritizing fairness alone may hinder task progress. Moreover, student grouping constitutes a high-stakes educational decision. Teachers need to understand why a particular grouping is recommended and must be able to adjust group compositions when situational factors change, such as emerging conflicts, role imbalances, or evolving task demands. If grouping mechanisms operate as opaque black boxes without traceable evidence or adjustable constraints, teacher trust and sustained classroom adoption are unlikely. Consequently, group formation should be formulated as a constrained multi-objective optimization problem that explicitly encodes fixed-size requirements, balances complementarity and equity, and outputs not only group assignments but

also auditable evidence rationales and interfaces for human intervention. This formulation enables algorithmic decision-making to support and complement pedagogical judgment, rather than replace it [9–11].

1.3 Scope

This doctoral project investigates alignment-aware intelligent group formation for teamwork in authentic primary school STEAM PBL classrooms. The scope is defined by a closed-loop perspective that links aligned pedagogical objectives with evidence-driven modeling and classroom-feasible decision making, as shown in Fig. 2. Also, this study focuses on grouping as an instructional decision embedded in project activities such as discussion, making, and presentation, rather than treating grouping as generic clustering or achievement-based tracking. The intended use is teacher-facing support. The outputs are designed to be directly actionable in classrooms and to remain editable through teacher preferences and rules so that contextual management needs can be respected without sacrificing methodological transparency.

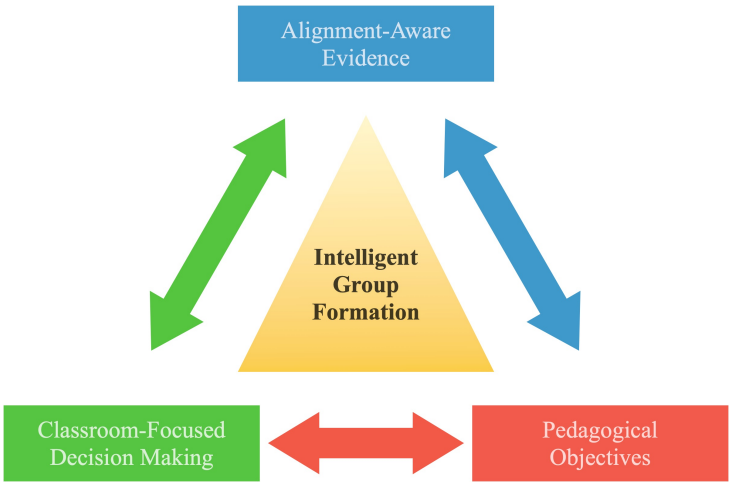


Figure 2: Intelligent group formation scope in STEAM PBL.

The evidence scope centers on heterogeneous classroom signals that can be indexed and traced back to their sources, while treating noise, partial missingness, and cross-modal misalignment as normal conditions. And it does not pursue peak recognition accuracy within any single modality. Instead, it examines how differences in evidence strength and trustworthiness influence grouping decisions, and how reliability-aware fusion can prevent recommendations from overreacting to noisy or weakly supported signals. The core mechanism operates on a chain of 'evidence reliability – relational structure – constrained decision-making'. Cross-modal alignment statistics and consistency modeling are used to quantify evidence reliability and to carry reliability information forward into subsequent modeling and decision steps. Collaboration is represented through student interaction graphs, and graph-based relational learning is employed to capture structural patterns and collaborative potential that are not accessible through static score-driven heuristics. Grouping is then formulated as a constrained multi-objective decision problem

that produces fixed-size groups while jointly considering within-group complementarity, between-group balance, and opportunity equity under classroom feasibility constraints. The resulting recommendations are accompanied by evidence-linked rationales that support traceability and auditability. The evaluation scope prioritizes operational metrics and controlled comparisons that are directly tied to classroom grouping requirements. The performance against representative baselines and the contributions of reliability modeling, relational learning, and constrained optimization are examined through ablation analyses.

To prevent scope creep, the study does not aim to develop new general-purpose multimodal foundation models, and it does not target end-to-end grouping directly from raw audio-video without intermediate evidence representations that can be inspected. It also does not cover settings in which the necessary heterogeneous evidence is infeasible to collect, and it does not prioritize grouping approaches that ignore interaction structure, omit classroom constraints, or lack an auditable rationale. Claims of generalization are treated conservatively, with applicability discussed primarily for settings that share comparable grade levels, PBL structures, and evidence availability, while broader transfer requires further empirical validation.

1.4 Research Questions

The decision to focus on STEAM PBL in primary school classrooms is grounded in the developmental characteristics of learners at this age. Primary school students experience rapid growth in social skills, collaborative awareness, and cognitive capacities. Consequently, grouping strategies can exert a strong influence on learning motivation, collaborative performance, and classroom experience. This study investigates evidence-driven collaborative group formation under authentic classroom constraints, where multimodal observations are heterogeneous and not uniformly reliable. The central research question is as follows:

How can alignment-aware multimodal evidence and graph-based relational modeling be leveraged to achieve fair, effective, and interpretable collaborative group in primary school STEAM PBL classroom?

Building on this core question, four specific research questions (RQs) were addressed:

RQ1 How can cross-modal alignment be quantified into alignment statistics that reflect the reliability of transcript evidence with respect to its aligned audio-video context, and how can such reliability signals be propagated to downstream learner representations under classroom noise and partial missingness?

RQ2 How can an interaction graph representation be constructed to capture students' collaborative relationships and complementary potential, while mitigating the influence of unreliable evidence and supporting structure-aware inference?

RQ3 How can an intelligent grouping mechanism be designed that integrates alignment-aware evidence signals with structure-aware graph modeling and produces fixed-size group assignments that balance within-group complementarity and between-group equity?

RQ4 How can objective functions and evaluation protocols be designed to ensure that grouping outcomes remain interpretable, traceable to evidence units, open to teacher intervention, and aligned with educational goals?

Based on these research questions, an alignment-aware methodological framework is proposed. First, heterogeneous classroom data are abstracted into traceable evidence units at appropriate granularities, enabling consistent aggregation and auditing across projects. Second, cross-modal alignment statistics are computed to quantify the reliability of utterance-level evidence and are used as reliability signals for weighting and filtering during feature construction and interaction modeling. Third, an interaction graph is constructed, and a graph-based relational model is employed to learn structure-aware learner representations that encode peer influence patterns and collaboration topology under reliability-aware edge semantics. Finally, a constrained multi-objective optimization module generates fixed-size grouping assignments that balance complementarity, collaboration potential, and fairness, producing evidence-based rationales that enable instructors to audit, interpret, and adjust grouping outcomes as needed.

1.5 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 1 introduces the research problem of intelligent group formation in primary-school STEAM PBL, formulates the research questions, and outlines the evidence-to-decision perspective under authentic classroom constraints.

Chapter 2 reviews the educational and computational foundations, including collaborative learning in childhood, learning analytics and educational digitalization, multimodal large language models, graph neural networks, and constrained multiobjective optimization.

Chapter 3 surveys state-of-the-art research on multimodal learning analytics, learner modeling, graph-based collaboration modeling, and algorithmic group formation, and it identifies gaps that motivate an alignment-aware and structure-aware yet classroom-feasible grouping framework.

Chapters 4 and 5 present two methodological pillars: graph-enhanced mutual learning and structure-aware graph attention modeling under heterogeneity, and discuss how their principles transfer to robust classroom analytics.

Chapter 6 documents the STEAM PBL study context and data collection, and specifies the evidence-chain dataset organization with traceable evidence units and quality/ethics considerations.

Chapter 7 presents the proposed end-to-end grouping methodology, integrating cross-modal alignment as reliability signals, interaction graph construction, GNN-based learner representations, and constrained multiobjective optimization with teacher-auditable rationales.

Chapter 8 reports the evaluation protocol, baselines, and results in terms of effectiveness, fairness, stability, and interpretability, including decision-level auditing and robustness analyses.

Chapter 9 concludes the thesis by summarizing contributions, discussing limitations, and outlining future work on scaling, richer evidence, longitudinal deployment, and teacher-AI co-design.

1.6 List of Publications

The work developed in this thesis has contributed to two academic publications, and these contributions closely follow the methodological and temporal progression of the project. The first publication, Graph-Enhanced Mutual Learning (GEML): A Graph-enhanced Pre-trained Language Model Framework for Text Classification via Mutual Learning, proposes a mutual-learning architecture that couples a pre-trained language model with a graph-based text classifier to better exploit document–word and document–document relations. Methodologically, this work provides the core idea of graph-enhanced representations and probability-distribution-level mutual learning, which is later adapted in this thesis to model students-students and student-multimodal features in a structured, graph-aware manner for student group formation. The second publication, Clustered Federated Spatio-Temporal Graph Attention Networks (CF-STGAT) for Skeleton-Based Action Recognition, introduces a clustered federated learning framework that leverages attention-derived spatio-temporal statistics for client clustering and inter-cluster model fusion. This work contributes the spatio-temporal graph attention backbone and the attention-guided clustering mechanism that are repurposed in this thesis to capture primary students' classroom behavior patterns and to inform graph-based grouping and role complementarity in STEAM PBL under realistic heterogeneity and privacy constraints. Building on these two technical pillars, the thesis further extends them from pre-trained large language models (LLM) and skeleton-based HAR settings to the education-specific scenario of MLLM-driven, graph-based intelligent grouping in primary school STEAM PBL, integrating multimodal representation, collaboration graph construction, and group optimization into a coherent, explainable, and teacher-interactive grouping framework.

Yu, T., Song, R., Pinto, S., Gomes, T., Tavares, A., & Xu, H. (2024). GEML: a graph-enhanced pre-trained language model framework for text classification via mutual learning. Applied Intelligence, 54(23), 12215-12229.

Yu, T., Pinto, S., Gomes, T., Tavares, A., & Xu, H. (2025). Clustered Federated Spatio-Temporal Graph Attention Networks for Skeleton-Based Action Recognition. Sensors, 25(23), 7277.

2

Background

This chapter presents the contextual background of this thesis. Section 2.1 introduces primary school STEAM PBL and the role of small-group collaboration, and reviews key learning and collaboration theories that explain why group composition matters for pupils' cognitive and socio-emotional development. Section 2.2 outlines multimodal learning analytics and the typical educational data modalities and summarizes the basic concepts of graph modeling, serving as the foundation for constructing student collaboration graphs. Finally, Section 2.3 reviews background on collaborative group formation and multi-objective optimization, highlighting the constraints for grouping in elementary school STEAM PBL classrooms.

2.1 STEAM PBL in Primary School

STEAM education breaks traditional subject boundaries and addresses authentic, real-world problems through integrated disciplinary learning. It advocates project-based approaches that highlight experiential and practical engagement [12, 13]. Its educational goal is to cultivate students' ability to apply interdisciplinary thinking to problem-solving, strengthen logical reasoning and creativity, and foster a broad range of 21st-century skills, thereby enabling learners to become innovative talents with a high level of STEAM literacy. At the international level, curriculum frameworks such as the Next Generation Science Standards (NGSS) in the United States resonate with STEAM principles by stressing the integration of scientific and engineering practices and the transfer of concepts across disciplines. This reflects the growing importance of cultivating integrated and innovative talent. As a global trend in the era of the knowledge economy, STEAM education is driving a transformation of educational paradigms and opening new pathways to improve students' scientific literacy and creative capacity. The Chinese education community has likewise responded to this development. As early as September 2015, the Ministry of Education explicitly encouraged primary and secondary schools, in national educational information technology policy documents, to explore STEAM education and other new forms of cross-disciplinary learning. Under the joint influence of policy guidance and classroom practice, STEAM has gradually been integrated into the basic education curriculum in China, reflecting a value orientation that seeks to develop students' core competencies and innovative spirit through disciplinary integration.

The core of PBL is to engage students in learning around a challenging problem or task, encouraging them to use a variety of resources in authentic contexts to solve a series of interrelated issues and produce concrete artifacts or solutions [14–16]. Throughout this process, students are expected to experience phases such as self-directed inquiry, teamwork, reflection, and improvement, and to apply disciplinary

knowledge in practice, thereby deepening their understanding of what they have learned. High-quality PBL is generally characterized by seven key elements: a challenging and driving problem, sustained inquiry over time, meaningful connections with real-world contexts, opportunities for student choice and voice, ongoing reflection, timely critique and revision, and public presentation of the final products. Research has shown that PBL helps students develop critical thinking, problem-solving, collaboration, communication, self-management, and other 21st-century skills, and promotes the transformation of knowledge into practical competence [17–19]. Compared with traditional teacher-centered, lecture-based instruction, it emphasizes real problems as the driving force for learning and encourages students to learn by doing and through collaboration. These features make PBL an important pathway for improving instructional quality and cultivating students' core competencies.

STEAM education and PBL are naturally aligned in both their aims and their methods, and their integration brings significant value and development opportunities to primary education. On the one hand, STEAM education seeks to solve real-world problems through interdisciplinary integration and to cultivate students' innovative awareness and ability to apply knowledge in a comprehensive way. In practice, it requires an appropriate pedagogical approach as its vehicle, and PBL provides precisely such an ideal pathway. PBL takes real-world problems as the starting point, encourages students to cross disciplinary boundaries in their inquiry, and guides them to integrate knowledge from different fields in the process of completing projects. Introducing PBL into STEAM courses uses project tasks as the core to achieve genuine integration of knowledge. For example, some researchers have proposed primary school STEAM curriculum models developed based on PBL, embedding multiple disciplines into the stages of project design, implementation, and evaluation, so that students experience the integrated use of knowledge from different subjects through hands-on practice [20–22]. Under this integrated model, students are confronted with complex topics that arise from everyday life and social needs, which they must solve by drawing on a wide range of knowledge and skills. In this way, they develop interdisciplinary thinking, a spirit of collaboration, and creativity. On the other hand, the use of PBL in STEAM contexts demonstrates distinctive instructional value. At the primary level, STEAM projects are usually closely related to students' lived experiences and are both engaging and challenging. Through PBL, students' motivation and level of participation increase significantly, and they strengthen their interest in science and technology as well as in the arts and humanities through firsthand exploration. Teachers can make use of the typical process, such as posing questions, inquiry, making, and sharing, to guide students through a complete inquiry cycle, connecting knowledge with real-life and fostering the ability to solve practical problems [23, 24]. In sum, the integration of STEAM education and PBL not only enriches the forms and content of the primary school curriculum but also effectively enhances students' overall competence and cultivates the innovative mindset and practical abilities required for future society. STEAM PBL in basic education has broad prospects for development and holds significant value for both research and practice.

2.1.1 Teamwork of Elementary Students in PBL

At the primary school stage, children are in a critical period of cognitive and social development, and these characteristics strongly influence their collaborative behavior. In terms of cognition, primary school students gradually move beyond the egocentric tendencies of the preschool years and enter the concrete operational stage described by Piaget [25, 26]. They begin to consider problems from the perspectives of others and to understand basic causality and rules. This improvement in perspective-taking and empathy makes them more willing to cooperate and to take their partners' feelings into account.

In the formation of collaborative learning groups, individual student characteristics serve as one of the important bases for grouping. Researchers have proposed that factors such as gender, personality type, social relationships, communication style, and ability profile should be considered together to achieve complementarity of strengths and coordination among group members [27]. For example, gender composition can influence group interaction. A highly homogeneous or severely imbalanced gender structure may lead to one group being ignored or the other dominating the discussion. In contrast, groups with a more balanced number of boys and girls tend to show more equitable interaction and smaller differences in academic outcomes. An empirical study on computer-supported collaborative learning further confirmed this point [28]. Groups with an equal number of boys and girls or composed entirely of girls achieved the best performance, whereas groups with one boy and three girls performed the worst. This suggests that extremely unbalanced gender combinations should be avoided in collaborative learning at the primary school level. Personality and communication style also shape how students participate in teamwork. Extroverted and introverted students tend to engage in different ways. Teachers often draw on their experience to place quiet and introverted students together with more active and talkative peers, and to establish appropriate rules for discussion so that each member has the opportunity to speak. Practice has shown that grouping decisions that take student personality into account can help improve the balance of interaction within the group [29]. Social relationships among students likewise affect the atmosphere of collaboration. Working with familiar friends can foster trust and communication, making members more willing to take responsibility and strive to achieve shared goals. However, some studies have reported that friend-based grouping does not always outperform grouping with unfamiliar peers. For instance, some research has found that groups of friends perform better at the beginning of a task but may relax their standards at later stages because of overfamiliarity, which leads to decreased efficiency [30]. Thus, the advantages and disadvantages of grouping based on social relationships remain inconclusive in the current literature and need to be judged in light of specific contexts. Overall, when forming collaborative groups among primary school students, it is advisable to consider multiple individual characteristics comprehensively, so as to ensure diversity and complementarity within groups and to reduce potential sources of conflict.

In actual classroom practice, teachers form collaborative groups on the basis of their accumulated classroom management experience and their understanding of students. Common strategies include deliberate

heterogeneous grouping, balanced ability distribution, and the avoidance of conflict-prone combinations [31]. First, many teachers follow the principle of "heterogeneity within groups and homogeneity between groups." In other words, each group is made as diverse as possible in terms of academic performance and gender, while the overall level of different groups is kept roughly equivalent. This heterogeneous grouping strategy is intended to enable higher-achieving students to support lower-achieving peers. Thus creating a cognitively balanced and mutually supportive environment, and at the same time preventing large gaps in performance between groups. Studies have shown that in groups of three or four students, heterogeneous academic composition often helps lower-achieving students receive more assistance, while higher-achieving students consolidate their own understanding through explaining content to others. Some fine-tuning is also reported: medium-level students sometimes participate more actively when grouped with peers at a similar level, so teachers may temporarily adopt homogeneous grouping to better address the needs of different achievement levels [32]. Second, teachers pay close attention to student compatibility and classroom management. Empirical investigations indicate that primary school teachers place the greatest emphasis on avoiding behavioral conflict when forming groups. They try to separate students who are likely to quarrel or cause discipline problems, and to distribute disruptive students across different groups. In contrast, teachers do not intentionally place close friends in the same group, and the promotion of friendship is regarded as the least important consideration. This practice aims to prevent group work from turning into off-task conversation or exclusive social cliques [33]. At the same time, teachers assign moderate importance to diversity in academic level, ensuring that each group includes students with different degrees of ability. Some research further suggests that although heterogeneous grouping may reduce the number of existing friends within each group, it can help lower the frequency of conflict, especially among girls [34]. This indicates that teachers seek to strike a balance between academic support and social harmony through grouping strategies. Finally, teachers' practical wisdom plays a crucial role in group formation. Experienced teachers take into account combinations of gender and personality, differences in academic level and interests, and whether students get along well with one another. In practice, many teachers refine group composition through continuous trial and adjustment, for example, by reorganizing or partially changing groups during the term in response to classroom feedback, to reach more effective collaborative pairings [35]. In sum, when teachers form collaborative groups, they aim to balance academic heterogeneity with interpersonal compatibility, and their strategies reflect both theoretical considerations and context-sensitive practical judgment.

Evaluating the effectiveness of collaborative groups among primary school students requires a process and outcome-based perspective. It should include multiple dimensions such as the quality of within-group interaction, the level of member participation, the quality of jointly produced outcomes, students' subjective learning experience, and the perceived fairness of collaboration [2, 27, 36].

- High-quality collaboration should be characterized by active and effective communication and negotiation. Observation indicators include whether group members engage in sufficient task-focused

discussion, whether meaningful questions and explanations are raised, and whether conflicts are appropriately resolved through communication. Studies have shown that the frequency of task-oriented interaction is closely related to learning performance. When students are required to share their ideas and solve open-ended problems together, more frequent and in-depth interaction often leads to more substantial knowledge construction and greater improvement in achievement. In contrast, if interaction is superficial and members work in isolation, it is difficult to achieve genuine collaborative gains.

- Collaborative learning emphasizes participation by all members. Therefore, evaluation should consider whether each student is actively engaged in group activities and whether any member is marginalized or remains "on the sidelines" of the discussion. Teachers can use classroom observation or video analysis to assess participation by examining indicators such as the number of times each student speaks and the proportion of task completion undertaken by each member. Groups in which all members participate tend to have a strong learning atmosphere, with higher levels of responsibility and motivation among students. In contrast, if only a few members dominate while most merely listen passively, the essence of collaborative learning is lost. For this reason, many teachers incorporate student self-evaluation and peer evaluation into the assessment process, encouraging group members to reflect on one another's contributions to quantify each student's level of involvement.

- The quality of the group's final product is a direct indicator of the effectiveness of collaborative learning. It reflects both the academic level and the creativity demonstrated in completing the task. Forms of evaluation may include group artifacts, reports, presentations, or test scores. In general, the quality of group output depends on the preceding processes of interaction and participation: a good collaboration process usually leads to high-quality outcomes. Research has also shown that interpersonal coordination within the group has a significant impact on final performance. Groups with positive interpersonal interaction tend to achieve higher task scores. In practice, teachers evaluate not only the overall quality of the group product but also often supplement it with individual assessment to prevent personal contributions from being hidden behind a collective grade. For example, some teachers ask students to engage in group discussion while requiring each member to submit an individual report. This helps ensure individual accountability and a more accurate measure of each student's understanding of the content.

- The value of collaborative learning lies not only in academic outcomes but also in its impact on students' social and emotional development. Evaluation should therefore include feedback on students' collaborative experiences, such as whether the group climate is harmonious, whether students enjoy the process of working together, and whether they gain confidence and a sense of belonging from collaboration. Some studies use the notion of "team satisfaction" to capture this dimension and show that positive interaction at both the task and interpersonal levels can

significantly enhance satisfaction with collaboration. In primary school settings, when students report feeling supported and treated as equals during group work, they tend to prefer this mode of learning, and their sense of classroom engagement and interest in learning also increase. Teachers can collect students' subjective experiences through questionnaires, interviews, or group reflection journals and use this information as an important reference for evaluating the effectiveness of collaborative learning.

- Fair collaboration means that every group member assumes responsibility and makes a genuine effort, and that no one consistently free-rides or is merely carried by others. This dimension is often regarded as one of the most challenging aspects of evaluating collaborative work. Teachers need to attend to whether tasks are distributed reasonably within the group and whether contributions are balanced. They must avoid situations in which one high-performing student secures a high grade for the entire group or an individual's efforts go unrecognized. In practice, many classrooms introduce peer evaluation mechanisms to strengthen fairness, inviting students to rate and comment on one another's contributions.

2.1.2 Educational Digitalization and AI Trends

In the context of the accelerating digital transformation of education, the classroom is no longer a "black box" of data, but has become a multi-source data environment that can be captured and analyzed. This shift has led to the emergence of large volumes of multimodal educational data that span both online and offline learning, including video, audio, text, behavioral logs, and assignments. For example, typical classroom data include video recordings of students in class, audio of classroom discussions, written homework, and spoken homework after class, as well as multimedia slides used during instruction. These data appear in multiple modalities such as images, speech, and text and are embedded throughout the entire teaching and learning process, providing an objective basis for instructional research and decision-making [37–39]. Driven by national strategies for educational digitalization, it has become a shared consensus that artificial intelligence and big data technologies should be used to collect and analyze classroom data in order to improve educational quality and promote educational equity. The Education Informatization 2.0 Action Plan issued in China in 2018 explicitly emphasizes the collection and analysis of big data and the deep integration of artificial intelligence into real classroom teaching, to achieve personalized instruction and teaching according to students' individual characteristics. In this sense, multimodal classroom data have become an essential resource for smart education, providing the data foundation for classroom diagnosis, learning behavior analysis, and instructional feedback. Research shows that, under the background of educational digitalization, introducing real-time classroom data observation and feedback helps identify weaknesses in the instructional process, optimize lesson design, and improve classroom effectiveness. With the support of artificial intelligence, classroom observation has entered a new stage of automated collection and intelligent analysis, enabling the objective recording of teachers' and students' verbal and nonverbal behaviors. Recent reviews indicate that commonly used data types

in digital classroom observation include facial expression data, verbal communication information, movement trajectories, body posture, and digital log data. Intelligent analysis of these multi-source classroom data can reveal key indicators such as student attention, interaction patterns, and emotional states. The acquisition and use of classroom data in the era of educational digitalization create unprecedented possibilities for gaining deeper insight into classroom practice, quantifying teaching processes, and supporting data-informed decision-making.

In the process of educational digitalization, large-scale pre-training models and various deep learning algorithms have shown great potential for understanding classroom data and optimizing student grouping. Pre-trained large language models such as BERT and GPT, which are based on the Transformer architecture, possess strong capabilities in language understanding and generation. They can be used for analyzing classroom textual content, interpreting dialogue, and generating instructional feedback. In recent years, general-purpose large models such as ChatGPT and Stable Diffusion, as well as domain-specific models for education, such as MathGPT for mathematical problem solving and "Ziyue," have further advanced the development of intelligent classroom applications. These models are able to perform semantic analysis on classroom discussions in both text and speech, automatically extract key concepts and points of confusion, and even generate instructional suggestions or answers to students' questions, thereby deepening and broadening classroom understanding. In addition, artificial intelligence technologies based on computer vision and speech recognition play an important role in the analysis of classroom video and audio data. For example, image recognition can be used to analyze students' facial expressions and body movements, while classroom recordings can be transcribed into text for further analysis. When combined with multimodal learning, pre-trained models make it possible to jointly understand video, audio, and text, offering new technical means for interpreting classroom situations. In summary, large-scale models and deep learning algorithms have given educational research the ability to "read" the classroom, from analyzing micro-level facial expressions and gestures, to modeling macro-level social learning networks, and further to predicting the evolving trajectories of learning behaviors. Together, these technologies provide powerful tools for understanding and optimizing classroom processes and establish an algorithmic foundation for intelligent grouping and personalized instruction. Thus, advanced AI algorithms are increasingly penetrating classroom settings, moving from classroom understanding to grouping optimization and providing solid technical support for more intelligent and efficient teaching.

The full use of instructional process data, combined with the empowering role of AI, makes it possible to enhance personalization, equity, and efficiency in education at the same time. For a long period, mass education, educational equity, and truly individualized instruction were regarded as goals that were difficult to achieve simultaneously. First, in terms of personalized learning, data-based approaches can provide customized learning paths and resource recommendations according to students' knowledge levels, learning styles, and interest preferences. With the support of learning behavior data and AI algorithms, teachers are able to carry out precise interventions tailored to different students. Second, regarding educational

equity and resource optimization, data and AI can help narrow gaps between regions and schools, enabling more students to share high-quality educational resources. In terms of instructional efficiency and teaching support, data-based intelligent tools greatly improve the effectiveness of teaching and educational management. On the one hand, AI can serve as a "digital assistant" for teachers by helping to automate tedious and repetitive tasks, allowing teachers to focus their energy on creative instruction. On the other hand, tools for analyzing teaching processes can collect classroom interaction data in real-time and generate classroom observation reports for teachers' reference. Overall, AI applications driven by data are empowering education in an all-around way. From assisting teachers in lesson preparation and classroom management to supporting personalized tutoring and learning assessment, every stage of the teaching and learning process is becoming more efficient. Educational policies and research in the field consistently argue that schools should actively embrace the use of AI technologies in instruction, learning, assessment, and administration, in order to reduce unnecessary burdens, increase efficiency, and improve the quality of education.

2.2 MLLM and GNN

With the continued development of digital instruction and intelligent technologies, an ecosystem of learning analytics has gradually taken shape in educational settings, one grounded in multimodal data and supported by intelligent models. Against this backdrop, the scope of large models in education is constantly expanding. Among them, MLLMs further strengthen the unified understanding and representation of multi-source information. At the same time, the increasingly complex interactions among students, learning resources, and tasks have drawn growing attention to GNNs, which can process graph-structured data in educational research. MLLMs demonstrate strong capabilities in semantic integration and cross-modal representation for multimodal data, whereas GNNs can effectively capture relational structure features that arise in the learning process. Together, they support educational data analysis from the two complementary dimensions of "content understanding" and "relationship modeling," forming an important technical foundation for current research in intelligent education. On this basis, the following sections will discuss three aspects in turn: multimodal data and MLLMs in education, graph-based models in educational applications, and the integration of MLLMs and GNNs.

2.2.1 Multimodal Data and MLLM in Education

With the rapid development of educational informatization and blended learning environments, data generated in students' learning processes increasingly exhibit multimodal characteristics. Beyond traditional textual artifacts such as reflections, quizzes, and assignments, contemporary classrooms produce rich streams of video (e.g., facial expressions, gestures, body posture), audio (e.g., prosody, turn-taking, speaking tempo), image data (e.g., photos of written work or artifacts), as well as clickstream logs and resource access traces from digital platforms. Taken together, these heterogeneous learning traces form a "digital

footprint" of the learner that reflects not only cognitive processes and learning strategies, but also engagement levels, emotional states, communication styles, and collaboration tendencies. However, extracting educationally meaningful features from such complex multimodal data has long been a central challenge in learning analytics and intelligent education. Traditional educational data mining approaches typically rely on hand-crafted statistical or rule-based features, and often focus on a single modality (e.g., only performance scores, log events, or questionnaire responses). As a result, they struggle to capture semantic relationships and behavioral patterns that span across modalities. Recent advances in deep learning and large-scale pre-training, especially MLLMs, provide a new technical pathway to address this limitation.

MLLMs are capable of processing multiple modalities such as text, images, and audio within a unified architecture, mapping them into a shared representation space and enabling cross-modal alignment and fusion. Through large-scale pre-training, these models can automatically extract high-dimensional, semantically coherent, and generalizable representations from raw classroom data, significantly reducing the need for manual feature engineering. In this process, textual input (e.g., reflections, reports, chat transcripts) can encode students' conceptual understanding and verbal style; video captures their attention, interaction patterns, and affective expressions; audio reflects fluency, participation, and emotional tone. By jointly modeling these signals, MLLMs support richer and more holistic student representations. In the context of student grouping, leveraging MLLMs to integrate multimodal evidence enables the system to more finely characterize students' cognitive profiles, interest structures, expressive styles, and collaboration tendencies. Compared to grouping strategies that rely only on grades or a small number of questionnaire dimensions, MLLM-based multimodal representations are more likely to reflect students' actual roles and behaviors in classroom interaction, thereby providing a stronger and more informative basis for subsequent relational modeling and grouping decisions.

2.2.2 Graph-based Model Applications in Education

Learning is inherently relational and social. On the one hand, students are embedded in networks of peer interaction through help-seeking, discussion, group work, and peer assessment. On the other hand, students are also linked to knowledge components, learning resources, and tasks through various dependency and alignment relations. Together, these elements form a complex graph-structured ecosystem in which students, resources, and tasks can be regarded as nodes, and their interactions and dependencies as edges. Such graph structures encode important information about group cohesion, knowledge pathways, learning dependencies, and social compatibility, which cannot be fully captured by treating samples as isolated instances. With the development of GNNs, an increasing number of educational studies have begun to construct graph structures over students, courses, concepts, or resources, in order to better capture relational dependencies in learning processes. GNNs adopt message passing and neighborhood aggregation mechanisms to iteratively integrate information from a node's neighbors while preserving its own attributes, thereby producing graph embeddings that reflect both individual properties and network structure. Based on this idea, graph-based models have shown clear advantages in a variety

of educational scenarios, including: Performance prediction and early warning, by modeling student–student or student–course relations to improve predictive accuracy; Analysis of student interaction networks, to identify key actors, marginal students, and interaction patterns for pedagogical intervention; Modeling student–knowledge and student–course graphs, enabling personalized recommendations, knowledge tracing, and learning path planning; Analysis of school social and collaboration networks, to uncover hidden cliques, potential conflicts, and opportunities for targeted support.

In student grouping tasks, the value of GNNs is particularly prominent. Grouping decisions depend not only on individual attributes (e.g., ability, personality, interests) but also on students' positions and neighborhoods in the relational network. Factors such as prior collaboration, interaction frequency, social compatibility, and peer acceptance can all be naturally modeled as edges or edge weights in a student graph. By performing graph embedding over such a structure, one can obtain student representations that encode both individual characteristics and collaboration potential. This, in turn, provides a powerful basis for structure-aware grouping decisions, enabling the formation of groups that better reflect actual interaction tendencies and the collaborative needs of classroom activities.

2.2.3 MLLM and GNN for Grouping

Although MLLMs excel at extracting rich multimodal features and GNNs are well-suited for modeling relational structure, relatively few studies have systematically combined the two and applied them specifically to student grouping. Many existing approaches to grouping either rely on limited, low-dimensional static features (e.g., grades, gender, personality scales) for clustering, or focus solely on graph partitioning based on interaction data. Consequently, they often fail to jointly leverage multimodal behavioral information and relational structure, which limits both the educational interpretability and the practical usefulness of the resulting group assignments. In practice, student grouping in classrooms is governed by a dense set of constraints and trade-offs, making it a complex and tightly constrained process, as detailed below:

1. *Data sparsity and incompleteness*–behavioral records may be incomplete for some students, different modalities may be missing to varying degrees, and sampling frequencies and qualities are often inconsistent.

2. *Feature heterogeneity and semantic misalignment*–textual, visual, and auditory data are heterogeneous in form and distribution, and cannot be directly compared in the raw space.

3. *Real-time requirements*–in practice, teachers often need usable grouping suggestions within a limited time, which constrains model complexity and inference latency.

4. *Explainability and controllability*–teachers need not only the grouping outcome, but also insights into why certain students are grouped, and the ability to adjust or override group assignments when necessary.

Traditional single-modality or simple feature-based clustering methods struggle to produce high-quality groupings under these combined constraints consistently. Against this backdrop, integrating MLLM and GNN offers a promising direction for constructing a grouping framework that jointly supports "content understanding" and "relational reasoning." In such a framework, the MLLM operates on multimodal classroom data to derive unified semantic representations of students, while the GNN operates on a student–student or student–task graph to perform structure-aware aggregation and inference. Concretely, one can first employ an MLLM to encode textual artifacts (e.g., questionnaires, written assignments, project reports, transcribed discussions) and associated visual or auditory cues into initial student embeddings. Then, based on collaboration history, interaction frequency, complementarity in abilities or personality, and other relevant factors, one constructs a relational graph over students. Applying a GNN over this graph refines the initial embeddings by integrating structural information, so that each student vector reflects both multimodal individual traits and their position in the collaboration network. Finally, clustering or grouping optimization (e.g., graph-based clustering with multi-objective constraints) is performed in this embedding space to produce groups that better balance intra-group complementarity and inter-group fairness.

This blended use of MLLM and GNN enhances the grouping system's capacity to handle complex multimodal data and relational information, leading to student representations with higher expressiveness and robustness. At the same time, the graph structure supports intuitive explanations: teachers can review individual student features and network-level relations, and revise group assignments when necessary. Such a framework thus provides a technically grounded path toward intelligent grouping that aligns with pedagogical goals and supports human–AI co-orchestration in real classrooms.

2.3 Collaborative Grouping and Multiobjective Optimization

In the context of the widespread adoption of collaborative learning and PBL, how to form high-quality groups according to instructional goals and student characteristics has become an important issue in both instructional design and learning analytics. Appropriate grouping is not only related to whether learning tasks can progress smoothly, but also directly affects peer support, knowledge construction, and learning motivation. Compared with traditional grouping methods that rely mainly on teacher experience or simple rules, data-based and intelligent grouping approaches can perform systematic optimization at a larger scale and under more complex constraints, gradually shifting grouping from an experience-based process to one that is goal-oriented. However, student grouping is essentially a combinatorial optimization problem with multiple objectives and multiple constraints. On the one hand, it is necessary to consider both within-group diversity and between-group balance across multiple dimensions such as academic performance, ability structure, interest preferences, and interaction relationships. On the other hand, grouping must also satisfy various hard constraints arising from teaching and management requirements, including group size, gender ratio, and disciplinary background. As a result, the grouping problem is difficult to solve with a single indicator or simple heuristic rules, and there is an urgent need to draw

on multiobjective optimization theory and intelligent algorithms to provide more systematic support for collaborative learning.

2.3.1 From Manual to Intelligent Grouping

In traditional teaching practice, student grouping has mainly relied on teachers' experiential judgment or simple random assignment. Common approaches include pairing high and low achievers based on test scores, grouping students according to seating or class list order, or manually adjusting groups based on the teacher's intuitive impressions of students' personalities and interpersonal relationships. These methods have advantages such as being easy to operate and quick to implement in small classes, but they also present clear limitations. First, they depend heavily on subjective teacher judgment and are difficult to make transparent or to evaluate quantitatively. Second, they typically draw on only a small amount of visible information, such as exam scores or classroom impressions, and thus cannot systematically integrate students' multimodal and multidimensional behavioral characteristics. Third, in large classes or in contexts where frequent regrouping is required, the workload and instability associated with manual grouping increase significantly. With the widespread adoption of learning management systems and educational technology platforms, students' process data and behavioral data are now recorded more systematically, providing a richer evidence base for group formation. On this foundation, some studies have begun to explore rule-based or simple data-driven grouping support systems, for example, using score ranges, interest tags, or questionnaire results to perform semi-automatic grouping, followed by manual fine-tuning by the teacher. These approaches partially reduce teachers' workload, but they still rely mainly on predefined rules and struggle to cope with complex situations involving multiple objectives and constraints [40, 41].

In recent years, driven by advances in machine learning, optimization algorithms, and large models, student grouping has gradually evolved toward intelligent grouping. Intelligent grouping no longer depends solely on fixed rules. Instead, student characteristics, interaction patterns, and instructional goals are formalized as computable features and objective functions, and methods such as clustering algorithms, graph-based models, evolutionary algorithms, and multi-objective optimization are introduced to solve the grouping problem. In this process, the teacher gradually shifts from being the direct executor who manually designs group compositions to becoming the designer who specifies goals and constraints, and the decision maker who reviews and intervenes in the grouping results. In this way, grouping decisions move toward a human–machine collaborative process.

2.3.2 Algorithms and Standards for Grouping

From an algorithmic perspective, student grouping can be regarded as a special type of partitioning or clustering task. The aim is to divide a given set of students into several groups under a set of constraints, in such a way that certain predefined objective functions are optimized. Existing research [42–44] have mainly followed several typical methodological paths. First, clustering approaches input student feature

vectors into clustering algorithms such as k-means, hierarchical clustering, or spectral clustering, and use similarity measures between samples to form several groups of "similar students." On this basis, instructors can choose either within-group similarity (homogeneous grouping) or within-group diversity (heterogeneous grouping) according to instructional goals. A common strategy is to perform similarity-based clustering first and then "break up" and reorganize students within each cluster in terms of ability or interest. Second, optimization and evolutionary algorithm approaches formulate grouping as a combinatorial optimization problem and use integer programming, heuristic search, genetic algorithms, particle swarm methods, and similar techniques to iteratively optimize grouping schemes. The advantage of these methods lies in their flexibility to define multiple objective functions and constraints and to search for approximately optimal solutions in a large solution space.

When evaluating grouping quality, general data analysis indicators can be used, such as within-group variance, between-group difference, diversity, and balance. In addition, criteria closely related to the educational context can be introduced, as follows:

1. *Alignment with instructional goals*–whether the grouping is conducive to the development of target competencies such as collaboration, creativity, or communication.

2. *Complementarity of abilities and roles*–whether each group contains a reasonable structure of abilities and potential for division of labor.

3. *Social and emotional factors*–whether students' social compatibility and sense of emotional safety are taken into account.

4. *Stability and sustainability*–the extent to which the grouping can be used across multiple tasks and its impact on longer-term collaborative relationships.

These algorithms and evaluation criteria together form a basic framework for intelligent grouping research and provide methodological support for the further introduction of multi-objective optimization and complex constraints.

2.3.3 Multiobjective Grouping and Constraints

In real classroom settings, student grouping usually serves multiple instructional goals simultaneously and is subject to various managerial and practical constraints. It is therefore more appropriate to model it as a multiobjective and multi-constraint optimization problem. On the one hand, teachers may wish to maintain a certain level of within-group heterogeneity, for example, diversity in ability, interest, or learning style, to promote the exchange of perspectives and peer support. On the other hand, they also hope to keep the overall level across groups relatively balanced and to avoid extreme situations in which high-achieving students are concentrated in a few groups while low-achieving students are clustered in others. In addition, teachers may attend to dimensions such as gender ratio, disciplinary background, role

distribution, and social network structure, which makes the grouping problem exhibit clear multiobjective characteristics.

With respect to constraints, grouping must satisfy hard constraints, such as upper and lower bounds on group size, limits on the total number of groups, special placement requirements for particular students, and time and resource constraints. At the same time, softer constraints related to instructional rhythm and system performance may also need to be considered, for example, generating an acceptable grouping plan within a limited time or supporting dynamic adjustment and partial regrouping during class. These factors make it difficult to adequately capture the grouping problem with a single objective or a simple scoring function.

Multiobjective optimization methods provide a systematic framework for addressing these issues. By formalizing different instructional goals as multiple objective functions, it is possible to use approaches such as weighted sums, hierarchical optimization, or evolutionary algorithms based on Pareto optimality to generate a set of alternative grouping solutions that embody different trade-offs among objectives. Teachers can then select from and intervene in these candidate solutions, creating an effective linkage between "algorithmically generated options" and "teacher-made final decisions." In this process, key questions include how to design objective functions and constraints that are educationally meaningful, how to maintain interpretability and controllability of grouping results during optimization, and how to reflect instructional preferences in the trade-offs among multiple objectives. These questions also provide conceptual and methodological interfaces for the subsequent introduction of richer student representations and more powerful structural modeling capabilities, such as methods that integrate MLLMs and GNNs.

3

State-of-the-Art

This chapter presents the state-of-the-art related to this doctoral work, covering group formation, learner modeling with MLLM, graph-based representations of learners and their interactions, and optimizing group formation. Section 3.1 reviews data-driven approaches to group formation in technology-enhanced collaborative learning, including classical heuristics, clustering-based methods, and metaheuristic algorithms, as well as their typical objectives and constraints. Then, section 3.2 surveys multimodal data modeling and the use of MLLMs in education, focusing on how diverse multimodal data can be encoded into unified learner representations. Section 3.3 introduces graph-based learner modeling, describing how student–student and student–knowledge relationships are captured and processed using GNNs and related techniques. Building on this, section 3.4 discusses multiobjective metaheuristics for optimizing group formation and other educational decision problems, highlighting how pedagogical criteria can be encoded as competing objectives within combinatorial search. Finally, Section 3.5 synthesizes these strands of literature and provides a gap analysis by contrasting existing solutions with the integrated multimodal, graph-based, and multiobjective framework proposed in this work.

3.1 Group Formation

Student grouping has long been regarded as a key component of collaborative learning design. It affects not only cognitive outcomes, but also peer support and affective experience. Several review studies have shown that grouping strategies can significantly affect the quality and outcomes of learning interactions and are therefore considered high-leverage decisions in the design of collaborative activities [45, 46]. Early work mainly relied on teachers' experience, achievement-based stratification, random grouping, or simple questionnaire rules. Typical practices include stratifying students according to test scores or prior knowledge, then forming groups through strong–weak pairing, or manually adjusting groups based on interest questionnaires and social relationships. These approaches are easy to implement and inexpensive, but they have apparent limitations, such as poor scalability, reliance on subjective judgment, and failure to systematically integrate learner characteristics. They also provide little support for frequent dynamic regrouping in real classrooms [47, 48].

With the development of learning analytics, a large amount of work has begun to model student grouping explicitly as a combinatorial optimization problem. Cruz and Isotani conducted a systematic mapping of grouping algorithms in collaborative learning contexts. They showed that existing methods can be broadly

classified as rule-based, clustering-based, optimization-based, and hybrid approaches. Additionally, commonly used student features (cognitive, behavioral, social, and affective) and evaluation criteria, such as within-group heterogeneity, between-group balance, and stability [46], are summarized. Maqtary further reviewed attributes and technical approaches for group formation and noted that many systems still rely on single-dimensional features, thereby failing to fully exploit the behavioral traces and multimodal data produced in online learning environments [45].

At the algorithmic level, particle swarm optimization (PSO) and genetic algorithms (GA) have been widely used for automatic grouping. Lin developed an automatic grouping system based on enhanced PSO, encoding student ability, learning style, and related characteristics into particle positions and using a fitness function to constrain both group size and within-group heterogeneity [49]. Other work has applied PSO to large-scale distance learning, constructing heterogeneous groups based on knowledge level and interest, and demonstrating the method's scalability in real online courses [50]. Another important line of research is based on GA. Moreno constructed fitness functions from multiple student features, including achievement, learning style, and social preference, and used GA to search for grouping schemes that satisfy various constraints [42]. In subsequent work, they proposed a hybrid framework that combines PSO, GA, and heuristic rules [51].

To address the dynamic nature of grouping, Srba and Bielikova treated dynamic grouping as a learning support strategy. They used learning process data to continuously evaluate the quality of within-group interaction and periodically reorganize the group structure based on student contributions and complementarity [52]. More recent work has proposed automatic grouping frameworks specifically for large-scale online courses, integrating students' prior knowledge, participation, and interaction behavior into the optimization objectives to improve the quality of online discussion [53]. Building on these ideas, multi-objective optimization has been introduced into the grouping problem. Miranda formulated grouping as a typical multiobjective combinatorial optimization task and proposed a multiobjective grouping framework that simultaneously optimizes within-group heterogeneity, between-group balance, and related indicators. Furthermore, the proposed framework uses multiobjective evolutionary algorithms to search for Pareto optimal solution sets [10]. Chen and Kuo proposed an optimized grouping method that explicitly models role distribution and problem-solving performance within a given problem context, thereby yielding groups more closely aligned with instructional goals [54]. The latest work further integrates knowledge tracing into the grouping process. For example, Li proposed a collaborative grouping framework based on deep knowledge tracing. They first diagnose students' knowledge states using an improved DKVMN model and then search for group compositions with complementary knowledge structures during optimization [55].

Overall, existing research on intelligent grouping has moved from experience-based and rule-based practices toward optimization models centered on group characteristics, and has gradually introduced multi-objective and dynamic grouping mechanisms. However, most work still relies mainly on limited structured features such as scores, questionnaires, and simple behavioral statistics, and rarely exploits multimodal

behavioral data or complex relational structures. At the same time, the coupling between grouping algorithms and underlying learner representations, such as deep feature embeddings or knowledge-tracing states, remains relatively loose, and a unified modeling framework remains lacking.

3.2 MLLMs in Education

Multimodal learning analytics (MMLA) provides a richer perspective on learner modeling than the traditional combination of grades and questionnaires. Blikstein first introduced the concept of multimodal learning analytics, emphasizing the use of video, interaction logs, and physiological signals to describe complex learning processes [56]. Later, Blikstein and Worsley discussed the integration of multimodal learning analytics and educational data mining, and proposed a framework that uses sensors and interaction logs to measure performance on complex tasks [37]. Di Mitri proposed a signal-to-knowledge model, which provides a layered conceptual framework for transforming low-level multimodal signals into high-level learning indicators [57]. Subsequent research has applied MMLA to engagement detection, recognition of affective and cognitive states, and analysis of collaborative processes. Typical data sources include classroom video and facial expressions, writing and programming traces, clickstream data, and eye gaze trajectories [38]. Ouhaichi provided a systematic mapping of multimodal learning analytics, summarizing trends in modality types, analytical methods, and application scenarios, and highlighting the need to link multimodal indicators to instructional decision-making more closely [58].

In parallel, the application of artificial intelligence in education has evolved from shallow to deep, and then to large models. A systematic review by Wang shows that recent AI in education research covers intelligent tutoring, automatic grading, learning prediction, resource recommendation, and other directions, with deep learning showing particular advantages for processing unstructured data such as text, images, and speech [59]. On this basis, LLMs are increasingly used for writing assistance, code tutoring, and conversational teaching. A systematic review of LLMs in education by Dong indicates that current applications are concentrated on text generation and conversational support. Also, learner modeling is still mainly based on textual interaction logs, with limited use of multimodal behavior [60]. More recently, large-scale multimodal foundation models, such as multimodal large language models, have created new opportunities for unified representation of multimodal learning behavior. Kuchemann discussed opportunities and challenges for multimodal foundation models in education, noting that such models can process language, images, audio, and video within a unified architecture and thus support richer representations of learning processes and cross-modal reasoning. At the same time, they also face challenges related to data privacy, bias, and interpretability [61]. Most current work still focuses on specific tasks, such as generating personalized feedback, detecting learning related emotions, or explaining learning resources. Systematic applications of large multimodal models for unified learner representation, together with downstream decision-making, such as grouping and intervention, remain relatively limited.

In summary, multimodal learning analytics and AI in education have already demonstrated the potential

to mine learning behavior from rich multimodal data. However, applications of LLMs and multimodal foundation models in education are still concentrated on content generation and interaction support. There is still a lack of systematic methods that place learner representation at the center and explicitly link multimodal behavior modeling to downstream decision tasks, especially to optimization and grouping.

3.3 Graph-Based Modeling

Graph structures are a natural way to represent many types of relationships in educational settings, such as student–student interaction graphs for discussion, collaboration, and peer assessment, student–course or student–resource bipartite graphs, knowledge concept graphs, and learning path graphs. With the development of GNNs, an increasing number of studies have sought to model learner attributes and multi-party relationships in a unified manner using graph-based methods. Zhou conducted a systematic review of GNN methods, summarizing mainstream architectures such as GCN and graph attention networks (GAT), and their applications to classification, link prediction, and clustering tasks, thereby providing a methodological foundation for graph modeling in education [62]. More recent reviews emphasize the extension of GNNs to many domains and the associated challenges, including over-smoothing, dynamic graph modeling, and scalability [63].

In education, GNNs have been used for a variety of tasks. Nakagawa proposed graph-based knowledge tracing (GKT), which models relationships among knowledge components as a graph and uses GNNs to update students' mastery of each knowledge component. GKT achieves more stable performance than traditional sequential knowledge tracing models [64]. Li proposed a multiview hypergraph neural network for academic performance prediction, encoding similarities among students across different categories of learning behavior as multiview hypergraphs and using hypergraph convolution to aggregate neighbor information and improve prediction accuracy [65]. Huang and Zeng proposed a dual graph neural network that models a student–activity interaction graph in one branch and a student attribute similarity graph in the other, and fuses local and global representations to improve performance prediction in online learning [66]. Recently, Hakkal systematically evaluated multiple GNN architectures on several educational datasets and showed that combining GNN-based graph embeddings with traditional ensemble models can significantly improve learner performance prediction [67]. These studies indicate that GNNs can effectively integrate individual features and interaction structures and can outperform traditional methods in learner modeling, knowledge tracing, and performance prediction. However, most work focuses on predictive tasks and only rarely addresses high-level instructional decisions such as grouping and intervention. Even when graph partitioning or community detection is used to form groups, the methods usually rely solely on structural information and lack deep integration with multimodal learner features.

Beyond predictive modeling, a growing but still limited body of work has begun to leverage graph-structured representations to support grouping decisions and instructional intervention. For example, some studies couple learner-state estimation with algorithmic grouping, where knowledge tracing outputs are used to

characterize learners' mastery profiles. Subsequent clustering or optimization forms collaboration groups that emphasize complementarity and equity [55]. Other work moves from individual diagnosis to group-level modeling by constructing group–student–item graphs and applying denoising GNN mechanisms to infer group cognitive states more robustly, thereby enabling teachers to identify which groups require targeted remediation and what conceptual gaps should be addressed [68]. In parallel, multimodal learning analytics has been used to monitor collaboration quality and trigger timely interventions, where audio–video cues, discourse signals, and interaction logs are fused to predict group collaboration satisfaction or engagement indicators that are actionable for teachers [69]. To improve interpretability for intervention planning, recent approaches further incorporate structure discovery, such as causal graph discovery over multi-view collaborative signals, to reveal behavior–outcome relations and provide more actionable explanations than black-box predictors [70]. Related efforts also propose structured group feedback artifacts, including group cognitive graphs that combine knowledge graphs with diagnostic estimates to visualize group-level strengths and weaknesses and guide both teacher scaffolding and students' self-regulation [71]. In large-scale online settings, interaction graphs and GNNs have additionally been used for early risk detection and proactive support, such as predicting dropout risk from learner–content–peer interaction subgraphs [72]. Collectively, these studies suggest that graph-based modeling can move beyond prediction toward decision support by linking relational structure to grouping and intervention. However, most existing approaches either rely on a single evidence source, treat multimodal signals without explicit reliability control, or stop short of producing fixed-size, constraint-satisfying grouping recommendations tightly coupled with auditable, evidence-linked rationales—leaving a clear opportunity for reliability-aware multimodal graph grouping frameworks.

3.4 Multiobjective Metaheuristics

As discussed earlier, student grouping is essentially a combinatorial optimization problem with multiple instructional goals and constraints. It is usually modeled as a search, under given class and group sizes, for grouping schemes that satisfy various heterogeneity and balance objectives [45, 46]. A large body of research has therefore focused on how to construct meaningful objective functions and constraints, and how to search efficiently for feasible solutions in a vast combinatorial space. In single objective optimization, early methods mostly used greedy or heuristic search with a single core indicator as the primary target, such as minimizing between-group variance in achievement or maximizing within-group heterogeneity. These methods are simple to implement, but they struggle to accommodate soft goals such as collaborative motivation or social preference. For this reason, swarm intelligence and evolutionary algorithms have been widely adopted. PSO-based approaches typically encode student features as particle positions, and design fitness functions that jointly describe within-group heterogeneity, between-group balance, and role distribution [49, 73]. GA-based methods explore the space of groupings through crossover and mutation, and fitness functions can flexibly combine multiple indicators based on the instructional context [42, 74].

Multiobjective optimization further improves the ability of grouping algorithms to represent complex instructional requirements. The multiobjective grouping framework proposed by Miranda: (1) models indicators, such as within-group heterogeneity and between-group balance, as separate objective functions, and (2) uses multiobjective evolutionary algorithms, such as Non-dominated Sorting Genetic Algorithms (NSGA) II, to search for a set of Pareto-optimal grouping solutions from which teachers can choose according to their preferences [10]. Some work on collaborative learning groupings incorporate deep knowledge tracing optimization strategies, using knowledge states from deep knowledge tracing models as optimization targets to promote group compositions with complementary knowledge structures [55]. Compared with grouping solely by scores, these methods place greater emphasis on learners' internal knowledge structures. Some studies view intelligent grouping from a more integrated perspective, combining clustering, matching, and evolutionary algorithms. The systematic mapping by Cruz and Isotani shows that hybrid methods have become a significant trend, for example, clustering students first to obtain coarse-grained segments, and then applying heuristic or evolutionary algorithms within each cluster for fine-grained grouping [46]. The hybrid framework proposed by Moreno combines heuristic rules, PSO, and GA, and improves solution quality and stability through stepwise repair and local search [51].

Despite this progress, most existing optimization frameworks still rely on limited structured features, and optimization objectives are primarily based on static scores, questionnaire results, and a few behavioral statistics. Although the work by Li already incorporates knowledge-state information from knowledge-tracing models into objective functions, learner modeling and grouping optimization remain only loosely coupled overall. Outputs from the modeling phase, such as estimated mastery levels, are usually treated as simple weights or constraints, and the richer internal representations of deep models, such as attention weights or graph embeddings, are rarely exploited. Moreover, few studies systematically consider the combined effect of graph-structural information (e.g., social or collaboration networks) and multimodal behavioral features in multiobjective optimization. And there is a lack of multiobjective grouping methods that operate under privacy and distribution constraints in federated environments.

3.5 Research Gap

The above work reveals several clear lines of development. First, in collaborative learning and intelligent grouping, many studies have formalized the grouping problem as a combinatorial optimization task and used clustering, PSO, GA, and multiobjective evolutionary algorithms to optimize indicators such as achievement balance, within-group heterogeneity, and role complementarity. Second, in learner modeling, multimodal learning analytics and AI in education have expanded data sources from grades and questionnaires to text, behavioral logs, and image or video data. LLMs and multimodal foundation models are being used for intelligent tutoring, automatic feedback, and emotion detection. Third, in structural modeling, GNNs have shown strong capabilities in knowledge tracing, performance prediction, and resource recommendation, effectively integrating learner attributes and interaction behavior through graph structures. Cross-domain work, such as GEML and CF-STGAT, demonstrates the potential of combining

graph models with large models and federated optimization. On this basis, the present study addresses three main gaps:

- *Lack of a unified learner representation that combines multimodal behavior and graph structure.* Existing work in multimodal learning analytics and large-scale models typically targets specific tasks, such as emotion recognition or automatic feedback. In contrast, GNN-based educational research focuses on tasks such as performance prediction and knowledge tracing. The two lines of work remain largely separate at the level of unified learner representation, and there is little research that systematically integrates multimodal behavior and relational structure within a single model.

- *Limited joint use of multimodal large models and GNNs in educational contexts.* Frameworks such as GEML, which couple PLMs with GNNs through mutual learning, have proven effective in general natural language processing, and CF-STGAT has shown the advantages of attention-driven clustering and fusion in federated settings. However, there is still no systematic attempt to transfer such graph-enhanced large models and attention-based federated clustering frameworks to learner modeling and collaborative group formation.

- *Insufficient integration of multimodal representation and structural information in multiobjective grouping under real instructional constraints.* Existing multiobjective grouping methods usually rely on a small set of structured features or single modality data. Even when knowledge tracing is used, only a few summary indicators are incorporated, and the richer internal representations of deep models, such as attention patterns and graph embeddings, are not fully exploited. Moreover, there is a lack of scalable multiobjective optimization methods that operate under privacy constraints and distributed data settings, and that jointly model graph structure and multimodal behavior.

To address these gaps, this study proposes a multiobjective grouping framework for collaborative learning. On the one hand, it draws on the idea of GEML to model students' multimodal behavioral features and relational structure with multimodal large models and graph neural networks, respectively, and to align these representations through a mutual learning mechanism. On the other hand, it incorporates the attention-driven federated clustering and lightweight cross-cluster fusion strategy of CF-STGAT to perform clustering and personalized updates of learner graph models in a federated environment. On this basis, the unified multimodal and graph-based learner representation is embedded in a multiobjective, multi-constraint optimization framework to generate collaborative grouping schemes that are interpretable, controllable, and adaptable to real classroom conditions.

4

Graph-enhanced Mutual Learning for Text Classification

This chapter introduces the GEML framework, which leverages mutual learning by combining GNN and pre-trained language models (PLM) for text classification. On the one hand, GEML provides a unified graph-enhanced baseline that systematically captures the synergy between graph structure modeling and sequence modeling. On the other hand, it proposes a general mutual learning paradigm in which knowledge is exchanged at the level of predictive distributions. In this way, GEML demonstrates how structured relational information, represented as graphs, can be integrated with highly expressive sequence models without modifying the internal architecture of existing PLMs. This design philosophy of being loosely coupled yet strongly coordinated will be extended in later chapters to more complex scenarios, including multimodal learner modeling and federated graph learning.

4.1 Motivation

Text classification is a fundamental task in the field of Natural Language Processing (NLP) and has been widely applied in sentiment analysis [75], information retrieval [76], intent classification [77], and more. Generally, text classification methods can be categorized into two main types: traditional methods and deep learning methods. The former techniques rely on hand-crafted features for text classification, such as sparse lexical features [78]. However, creating these features manually can be time-consuming and inefficient. Other traditional features include topic models [79] and TF-IDF [80]. These features are then used in statistics-based classifiers like Naive Bayes [81] and Support Vector Machine [82]. The latter methods avoid the need for hand-crafted features by automatically learning dense vector representations from the text. These methods are based on deep neural networks, such as Convolutional Neural Networks (CNN) [83] and Recurrent Neural Networks (RNN) [84]. And they are limited by often focusing solely on the text's context and cannot capture information sharing between distant, discontinuous words.

With the rapid development of representation learning, PLMs, such as BERT [85], RoBERTa [86], and their variants, have become the de facto backbone for text classification. By leveraging large-scale unsupervised pre-training and the self-attention mechanism, PLMs can capture rich contextual dependencies and significantly outperform earlier CNN-based and RNN-based architectures on many benchmarks. Nevertheless, most PLM-based methods still treat documents as independent and identically distributed (i.i.d.) instances and primarily model intra-sentence or intra-document context. As a result, they are less effective

at exploiting corpus-level relational information, such as global topic structure, inter-document similarity, or higher-order word co-occurrence patterns. In parallel, graph-based methods such as TextGCN construct corpus-level heterogeneous graphs in which both words and documents serve as nodes, and edges encode word–word or document–word relations [87]. Typical designs connect words that frequently co-occur within a sliding window and link documents to informative words using weighting schemes such as TF-IDF. GNN is then applied to this heterogeneous graph to propagate information and perform document-node classification. Compared with purely sequential models, graph-based methods are better at modeling long-range and global dependencies, and can naturally support label propagation across the corpus. However, they usually rely on bag-of-words-style node features and ignore fine-grained token order and contextual nuances within each document, making them complementary rather than competitive to PLMs.

Therefore, recent research explores combining PLMs and GNNs for text classification. A straightforward strategy is to treat PLM-derived sentence embeddings as node features in a text graph, or to inject graph-derived representations into the PLM encoder [88]. While such tightly coupled architectures can improve performance, they often require non-trivial modifications to the PLM's internal structure, complicating implementation, reducing modularity, and limiting the flexibility to swap different PLM backbones. In this work, we adopt a loosely coupled yet strongly cooperative framework. Instead of forcing architectural fusion, we keep the PLM and the graph model structurally independent, and let them interact at the probability-distribution level through a mutual learning mechanism [89]. Concretely, a PLM-based classifier focuses on capturing rich local and contextual semantics, while a GNN-based classifier operates on a corpus-level text graph to encode global relational structure. Their softened output distributions are then aligned via knowledge exchange, allowing each model to act as both a teacher and a student for the other. This framework not only boosts text classification performance but also provides a generic blueprint for combining structured graph information with powerful sequence models without invasive architectural changes.

GEML is designed to address these issues with three goals. First, preserve the architectural independence of the graph model and the PLM, enabling plug-and-play usage of different PLMs. Second, allow the two models teaching each other via mutual learning on probability distributions, instead of hard-wiring their internal representations. Third, improve robustness and generalization by smoothing and regularizing predictions through cross-model consistency.

Beyond the text classification scenario considered in this chapter, the same design principles naturally extend to more complex multimodal and graph-enhanced settings. In particular, when moving to educational applications that involve MLLMs and graph-based learner modeling, GEML offers a transferable paradigm: different modalities or subsystems (e.g., textual feedback, behavioral graphs, and physiological signals) can be modeled by specialized networks, and their knowledge can be integrated through mutual learning at the distribution level. In this sense, GEML serves as a centralized, graph-enhanced baseline and methodological bridge between classical text classification and the multimodal learner modeling tasks

explored in later chapters.

4.2 GEML Framework

As illustrated in Fig. 3, GEML contains two structurally independent yet cooperatively trained channels: a graph-based classifier operating on a heterogeneous word–document graph and a PLM-based classifier operating on token sequences. In the graph channel, each document is modeled as a document node D_i, and each vocabulary item is modeled as a word node W_i; edges encode corpus-level statistical relations with type-specific weighting schemes. Concretely, the document-word edges (D_i, W_j) are weighted by PPMI to reflect the strength of document–term association, while the word–word edges (W_i, W_j) are weighted by TF-IDF to capture global co-occurrence and importance relations among words. A multi-layer GCN propagates information over this graph to produce document-level logits, whereas the PLM branch produces logits from contextualized representations via a standard classification head. The two branches remain architecturally decoupled, but are coupled through mutual learning by aligning their softened output distributions and being trained with asymmetrical optimization. The asymmetrical optimization coordinates mini-batch updates in the PLM channel with full-batch propagation in the graph channel, where uncertainty-weighted loss composition balances supervised objectives and cross-model consistency.

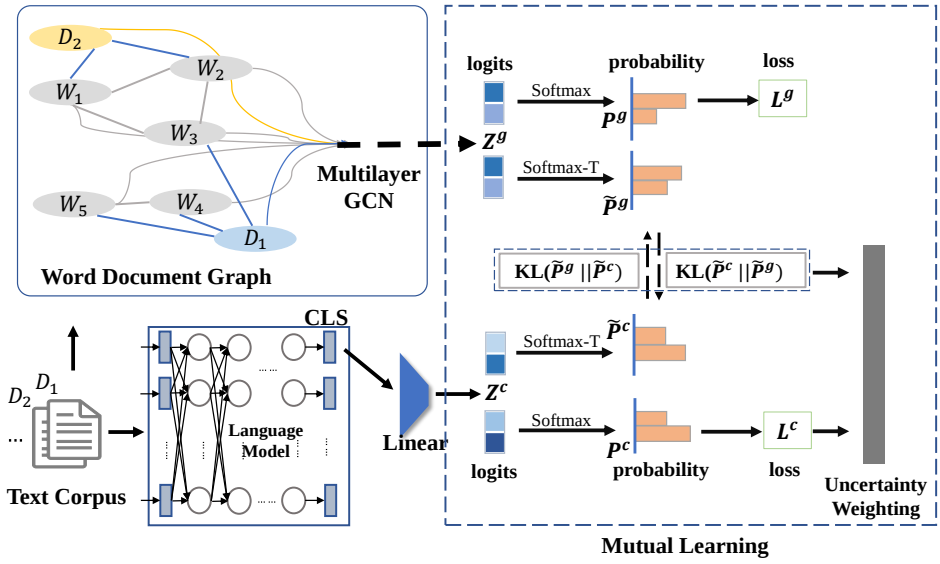


Figure 3: Model framework of GEML.

4.2.1 Problem Setting

Let $\mathcal{D}$ be a corpus of documents and C the set of labels. Text classification can be viewed as learning a mapping,

$$\zeta : x \in \mathcal{D} \rightarrow y \in C, \tag{4.1}$$

where each document x is associated with a label or label distribution y. In GEML, this mapping is realized by two parallel components: a graph-based model M_g that operates on a heterogeneous word–document graph and a context-based model M_c, instantiated as a pre-trained language model (PLM) equipped with a standard classification head. Given an input instance, both M_g and M_c produce logits over the label space C, which are subsequently transformed into class probabilities (e.g., via a softmax function).

The graph-based channel adopts a TextGCN-style architecture to exploit corpus-level structural information. Specifically, we construct a heterogeneous graph $G = (\mathcal{V}, \mathcal{E})$ whose node set $\mathcal{V}$ comprises both document nodes and word nodes, while the edge set $\mathcal{E}$ encodes relations such as word co-occurrence and word–document association. Edge weights are assigned according to relation type: (1) word–word edges are weighted by Positive Point-wise Mutual Information (PPMI) to capture co-occurrence statistics, (2) document–word edges are weighted by TF–IDF to reflect term importance within documents, and (3) self-loops are included to improve stability during graph convolution.

Each node is initialized with a one-hot feature vector and propagated through a two-layer Graph Convolutional Network (GCN). The second layer maps node representations into the label space, and the rows corresponding to document nodes constitute the logits matrix Z_g for classification. This design offers two key advantages: it exposes the PLM to corpus-level relational cues that are unavailable under independent document processing, and it is model-agnostic in the sense that TextGCN can be replaced by any transductive GNN that operates on the same corpus graph.

The PLM channel M_c follows the standard fine-tuning paradigm. Each document is tokenized and fed into a PLM such as DistilBERT, BERT, or RoBERTa. Typically, the representation corresponding to the `[CLS]` token (or an equivalent pooled embedding) is passed to a task-specific linear classifier to produce logits Z_c over the label set C. Crucially, GEML does not modify the internal architecture of the PLM. The integration is entirely carried out at the level of output probabilities and losses, preserving the PLM as a reusable backbone that can later be swapped or upgraded, for instance, by larger language models or domain-adapted variants.

4.2.2 Mutual Learning via Probability Distributions

The core of GEML is a mutual learning mechanism between M_g and M_c. Rather than treating the graph model as a fixed teacher and the PLM as a student, the two models learn from each other in a symmetric manner. Let Z_g and Z_c denote the logits produced by M_g and M_c, respectively. For each model, we construct two probability distributions: a standard softmax distribution with unit temperature, denoted P^g and P^c, which is used in the supervised cross-entropy objective, and a temperature-smoothed distribution with $\tau > 1$, denoted $\tilde{P}^g$ and $\tilde{P}^c$, which serves as the softened target for mutual learning.

The supervised loss for each model is the usual cross-entropy between the ground truth and its own

prediction:

$$L^g_{\text{sup}} = \text{CE}(Y, P^g), \qquad L^c_{\text{sup}} = \text{CE}(Y, P^c). \tag{4.2}$$

To encourage the two models to approximate each other's predictive behaviour, GEML adds bidirectional KL divergence terms between the smoothed distributions:

$$L^g_{\text{KL}} = \text{KL}(\tilde{P}^g \,||\, \tilde{P}^c), \qquad L^c_{\text{KL}} = \text{KL}(\tilde{P}^c \,||\, \tilde{P}^g). \tag{4.3}$$

Intuitively, each model is encouraged to move its probability mass closer to that of its peer, while the temperature τ controls the smoothness of the distributions and thereby how much information about relative class preferences is preserved. Experiments show that moderate temperatures ($\tau \approx 2$–3) tend to produce the best knowledge transfer.

4.2.3 Asymmetrical Learning

In practice, PLMs and graph models exhibit different convergence behaviours. The PLM, being pre-trained on large corpora, tends to converge quickly, whereas the graph model, initialized randomly and optimized in a fully transductive manner, may converge more slowly. If both models are updated at the same pace, the slower one can bottleneck mutual learning, and early noisy predictions can degrade the other model.

GEML introduces an asymmetrical learning strategy inspired by asymmetric graph representation learning. Concretely, during each PLM epoch, the graph model M_g is updated multiple times. A hyperparameter μ controls how many graph updates occur per PLM update. This allows the graph model to "catch up" and provide more meaningful probability distributions for mutual learning. A small warm-up phase (typically one epoch for each model) is also employed so that both models form reasonable initial predictions before they begin to learn from each other.

4.2.4 Uncertainty-Weighted Loss Composition

The joint loss for each model combines a supervised term and a mutual learning term. Rather than manually adjusting the weights between these terms, GEML adopts an uncertainty-weighting strategy originally proposed for multi-task learning. Each task (supervised prediction and mutual alignment) is associated with a learnable noise parameter σ, and the combined loss takes the form

$$L = \frac{1}{2\sigma_1^2} L_{\text{sup}} + \frac{1}{2\sigma_2^2} L_{\text{KL}} + \log(\sigma_1 \sigma_2). \tag{4.4}$$

This formulation allows the model to automatically adjust the emphasis placed on the supervised signal versus the mutual learning signal as training progresses. Empirically, introducing uncertainty weighting yields consistent but moderate improvements and reduces the need for manual loss-weight tuning.

4.2.5 Optimization with Document–Node Bijection

A practical challenge arises from the mismatch between the two channels' optimization regimes. The graph model operates on the entire corpus graph and is naturally trained in full-batch mode, whereas the PLM must be trained with mini-batches due to memory constraints. To compute the KL divergence between their distributions on the same documents, we must align samples across these two regimes. GEML resolves this mismatch between mini-batch PLM optimization and full-batch graph propagation by maintaining a *document–node bijection* (Fig. 4, *Document–Node Bijection*). Concretely, each document in the corpus is assigned a unique node index in the graph and a unique instance index in the PLM dataset; the bijection function $\mathrm{Bijection}(\cdot)$ then establishes a one-to-one correspondence between these two index spaces. In the illustrative case of six documents partitioned into three PLM mini-batches, the mapping aligns every mini-batch sample with its corresponding graph document node (word nodes in TextGCN are omitted for clarity); for example, $\mathrm{Bijection}(T_3) = N_5$ and $\mathrm{Bijection}(N_1) = T_2$. During training, when updating M_c, the KL divergence is computed between PLM predictions on a mini-batch and graph predictions on the matched document nodes obtained via $\mathrm{Bijection}(\cdot)$. Conversely, when updating M_g, PLM logits produced across all mini-batches are accumulated, stacked, and reordered according to graph node indices, so that the mutual-learning signal is always defined over correctly aligned samples. This bijection-based alignment guarantees consistent cross-model distillation despite the different batching strategies used by the PLM and the graph channel.

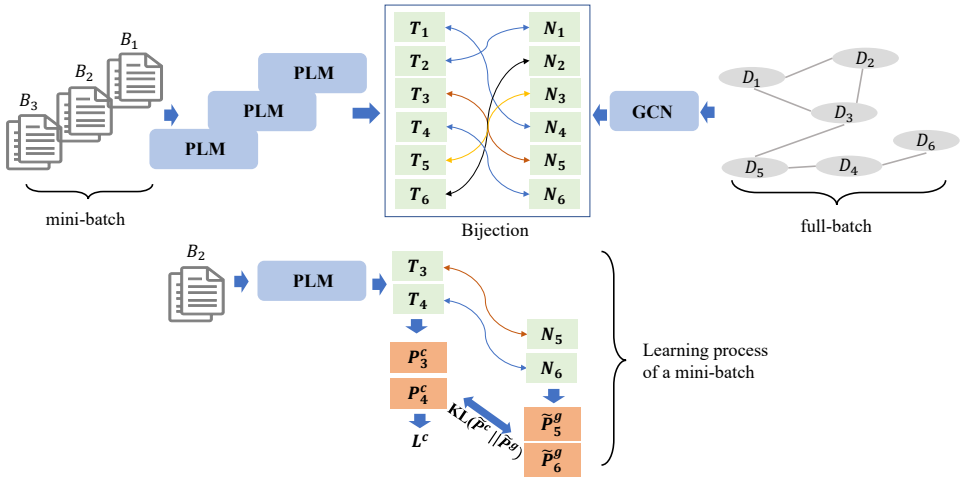


Figure 4: A bijection example.

4.3 Experimental Evaluation

This section summarizes the empirical evaluation of GEML on standard benchmarks, highlighting the benefits of mutual learning over strong graph-based and PLM baselines and relating the observed behaviours to the design choices introduced above.

4.3.1 Datasets

Datasets	**Train**	**Test**	**Classes**	Average length	Range of length
mr	7,108	3,554	2	18	1–52
SST	8,117	3,169	5	15	1–51
SearchSnippets	8,636	3,704	8	16	2–37
R8	5,485	2,189	8	101	1–963
R52	6,532	2,568	52	108	2–1038
Ohsumed	3,357	4,043	32	134	5–475

Table 1: Statistics of datasets.

GEML is evaluated on six widely used text classification datasets covering sentiment analysis and topic classification:

- **MR** A binary movie review sentiment dataset, where each instance is a single-sentence review labeled as positive or negative.
- **SST** A fine-grained sentiment dataset derived from movie reviews, annotated with five sentiment categories, making it substantially more challenging than MR.
- **R8** and **R52** Two subsets of the Reuters-21578 newswire corpus with 8 and 52 categories respectively, focusing on topic classification.
- **Ohsumed** A medical abstract dataset with 32 categories, derived from the MEDLINE database and centered on cardiovascular diseases.
- **SearchSnippets** Short text snippets returned by web search engines, organized into eight domains, including business, computer, education, and health. These datasets differ significantly in size, number of classes, domain, and average document length, providing a diverse testbed for evaluating the robustness of GEML.

4.3.2 Baselines and Implementation Details

The baselines fall into four groups:

- **Traditional neural architectures** FastText, TextCNN, and SWEM, which rely on shallow word embeddings with pooling or convolutional layers;
- **Graph-based models** S-LSTM, TextGCN, and TextING / TextING-M, which treat text classification as graph or sequence classification with explicit structural modeling;
- **PLMs** DistilBERT, BERT, and RoBERTa, fine-tuned with a standard classification head;
- **Graph-enhanced PLMs** VGCN-BERT and BERTGCN (and their DistilBERT / RoBERTa variants),

which combine graph models and PLMs through feature concatenation or probability interpolation. GEML is instantiated with TextGCN as M_g and with different PLMs as M_c, resulting in variants such as GEML+DistilBERT, GEML+BERT, and GEML+RoBERTa.

The graph channel uses a two-layer GCN with a hidden dimension of 300, trained with Adam. The learning rates for the PLM and the graph model are set to $1\mathrm{e}{-5}$ and 0.01, respectively. The number of epochs is fixed across experiments, and the hyperparameters μ (asymmetrical learning updates) and τ (temperature) are tuned on each dataset, typically with μ ranging from 1 to 5 and τ from 2 to 4. All models are implemented in PyTorch. For fair comparison, each configuration is run multiple times with different random seeds, and the mean and standard deviation of performance metrics are reported.

	mr	R8	R52	Ohsumed	Snippets	SST	Average
FastText	72.17±0.21	96.13±0.19	92.81±0.26	57.70±0.44	88.56±0.28	43.61±0.36	75.16
TextCNN	77.75±0.41	95.71±0.45	91.74±0.15	66.18±0.47	89.52±0.12	42.51±0.16	77.24
SWEM	76.65±0.34	95.32±0.27	92.94±0.31	63.62±0.40	87.36±0.28	41.62±0.11	76.25
S-LSTM	77.75±0.48	97.67±0.11	94.92±0.34	65.47±0.21	87.54±0.30	42.46±0.18	77.64
TextGCN	76.78±0.18	97.07±0.39	93.60±0.29	68.40±0.25	87.72±0.23	40.16±0.31	77.29
TextING	79.82±0.36	98.04±0.45	95.48±0.25	70.42±0.46	87.03±0.29	46.29±0.13	79.51
TextING-M	80.19±0.13	98.23±0.31	95.68±0.14	70.84±0.34	88.53±0.14	47.25±20	80.12
DistilBERT	84.81±0.14	97.49±0.21	96.34±0.13	69.70±0.18	91.22±0.25	50.36±0.24	81.65
DistilBERTGCN	85.03±0.27	97.91±0.13	96.45±0.10	70.18±0.08	91.38±0.15	50.66±0.12	81.94
GEML+DistilBERT	**85.58±0.35**	**98.31±0.30**	**97.04±0.29**	**71.75±0.21**	**91.87±0.34**	**50.80±0.27**	**82.56**
BERT	85.59±0.15	98.08±0.19	96.46±0.29	71.38±0.21	91.52±0.17	51.31±0.10	82.39
BERTGCN	86.18±0.24	98.13±0.16	96.60±0.23	72.80±0.18	91.79±0.13	51.64±0.06	82.86
VGCN-BERT	86.30±0.15	98.17±0.20	96.11±0.19	71.49±0.11	91.71±0.17	51.33±0.21	82.52
GEML+BERT	**86.72±0.29**	**98.45±0.23**	**96.96±0.18**	**73.18±0.27**	**92.31±0.24**	**52.13±0.19**	**83.29**
RoBERTa	89.50±0.08	98.17±0.13	96.20±0.21	72.99±0.16	90.63±0.10	55.31±0.11	83.80
RoBERTaGCN	89.70±0.15	98.31±0.08	96.10±0.23	72.80±0.20	91.41±0.14	54.68±0.07	83.83
GEML+RoBERTa	**89.81±0.21**	**98.49±0.30**	**96.92±0.22**	**73.33±0.32**	**91.44±0.27**	**56.23±0.16**	**84.37**

Table 2: Accuracy results on different datasets (%).

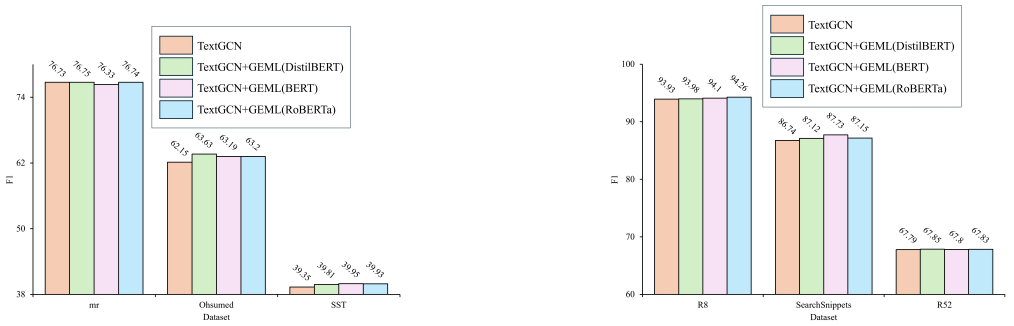
	mr	R8	R52	Ohsumed	Snippets	SST	Average
FastText	75.61±0.37	90.57±0.42	57.98±0.40	45.71±0.36	87.03±0.28	38.42±0.29	65.89
TextCNN	77.39±0.33	93.56±0.44	60.80±0.32	56.08±0.34	88.27±0.49	41.03±0.22	69.52
SWEM	75.51±0.39	89.29±0.41	48.27±0.22	49.81±0.25	87.09±0.47	39.20±0.35	64.86
S-LSTM	76.96±0.27	93.94±0.44	73.33±0.25	61.39±0.32	86.63±0.31	39.65±0.24	71.89
TextGCN	76.73±0.28	93.93±0.37	67.79±0.24	62.15±0.39	86.74±0.47	39.35±0.22	71.12
TextING-M	80.02±0.17	93.62±0.22	73.38±0.15	61.37±0.31	87.26±0.20	42.26±0.26	72.99
DistilBERT	84.81±0.14	94.42±0.20	83.79±0.18	59.48±0.12	89.25±0.27	49.17±0.13	76.82
DistilBERTGCN	85.06±0.18	94.67±0.17	84.33±0.23	61.29±0.24	90.42±0.22	49.23±0.14	77.50
GEML+DistilBERT	**85.69±0.31**	**94.83±0.37**	**85.59±0.27**	**62.93±0.24**	**91.09±0.32**	**49.70±0.21**	**78.31**
BERT	85.62±0.22	95.05±0.19	85.67±0.17	62.24±0.10	90.22±0.12	49.86±0.18	78.11
BERTGCN	86.25±0.25	95.76±0.14	86.04±0.27	63.07±0.11	90.30±0.19	50.54±0.24	78.66
VGCN-BERT	86.10±0.17	95.53±0.29	86.17±0.23	63.28±0.18	90.44±0.15	50.32±0.16	78.64
GEML+BERT	**86.71±0.35**	**96.59±0.30**	**86.78±0.26**	**64.13±0.37**	**91.05±0.32**	**51.45±0.22**	**79.45**
RoBERTa	89.71±0.18	95.01±0.23	85.11±0.17	62.80±0.09	90.06±0.13	53.02±0.20	79.29
RoBERTaGCN	89.50±0.21	95.44±0.15	86.32±0.29	63.39±0.13	90.78±0.24	53.55±0.12	79.83
GEML+RoBERTa	**89.80±0.25**	**96.08±0.29**	**86.88±0.32**	**64.02±0.35**	**91.13±0.17**	**53.64±0.21**	**80.26**

Table 3: Macro-average F1 results on different datasets (%).

4.3.3 Overall Results

Table 2 and Table 3 report accuracy from the original GEML paper and macro-average F1 scores for all datasets and baselines, respectively. The key observations can be summarized as follows:

- **Graph models vs. PLMs** Pure graph-based methods (TextGCN, TextING, TextING-M) consistently outperform traditional CNN/RNN architectures, confirming the value of structural information. However, they still trail PLMs, especially on more complex datasets. This motivates using PLMs as the main performance backbone.
- **GEML vs. PLMs** For every PLM backbone (DistilBERT, BERT, RoBERTa), adding GEML yields consistent improvements in both accuracy and macro-F1 across the six datasets. The gains are particularly notable on more challenging datasets such as SST and Ohsumed, where the additional structural information and regularization provided by the graph model help the PLM avoid overconfidence and better handle rare or ambiguous patterns.
- **GEML vs. other graph-enhanced PLMs** GEML outperforms VGCN-BERT and BERTGCN variants on average. This indicates that probability-level mutual learning can be more effective than simple feature concatenation or one-way interpolation, as it enables bidirectional knowledge transfer and better regularizes the learning dynamics.



(a) F1 on mr, Ohsumed and SST

(b) F1 on R8, R52 and SearchSnippets

Figure 5: Mutual learning results.

Interestingly, even though TextGCN is clearly weaker than PLMs when used individually, the PLM still benefits substantially from learning against it. This aligns with the intuition that mutual learning leverages complementary inductive biases: the PLM contributes strong contextual language understanding, while the graph model contributes corpus-level relational structure and smooths decision boundaries. Although the thesis primarily views GEML as a vehicle for enhancing PLMs, the graph model also improves through mutual learning. Figure 5 in the original paper shows that TextGCN's macro-F1 increases when learning jointly with different PLMs, with larger gains when paired with stronger PLMs such as RoBERTa. The magnitude of improvement is modest compared to that of the PLM, partly because TextGCN's representational capacity and hyperparameters were not aggressively tuned. Nevertheless, this result is conceptually

important: weaker but structurally rich models can be improved through collaboration with stronger sequence models, demonstrating the generality of the mutual learning paradigm. Several ablation studies shed light on the role of GEML's design components:

- **Asymmetrical Learning (μ)** Varying the number of TextGCN updates per PLM epoch shows that appropriate asymmetry can significantly affect performance, especially on larger datasets such as R8, R52, and Ohsumed. When μ is too small, the graph model lags and provides less informative distributions; when μ is too large, training becomes inefficient. The best μ depends on dataset complexity but typically lies between 3 and 4.

- **Temperature (τ)** Experiments varying τ indicate that temperatures greater than one consistently outperform the standard softmax case ($\tau = 1$). This confirms that smoother probability distributions carry richer information for mutual learning and act similarly to label smoothing, improving generalization. Many datasets achieve the best results at $\tau \approx 3$.

- **Uncertainty Weighting** Removing uncertainty weighting and replacing it with a simple sum of losses causes small but consistent performance drops across datasets and models. This suggests that automatically balancing the relative importance of supervised and mutual learning components is beneficial and reduces the effort of manual tuning.

- **Loss Dynamics** Visualizing the training loss of BERT with and without GEML reveals that GEML produces smoother loss trajectories and converges to regions that generalize better, even when the final training loss is not significantly lower. This supports the view that mutual learning helps escape sharp minima and promotes wider, more robust optima.

4.4 Discussion and Connection to the Thesis

Beyond its standalone contributions to text classification, GEML plays a strategic role in the overall narrative of this thesis. It embodies a design pattern: graph-enhanced mutual learning, which is later generalized to multimodal learner modeling and federated graph learning. In the experiments, GEML operates on generic sentiment and topic classification datasets. However, the underlying modeling idea is directly transferable to educational settings. There, one can construct heterogeneous graphs whose nodes represent learners, learning resources, activities, or textual artefacts, with edges capturing interactions such as co-participation, prerequisite relations, or temporal proximity. A PLM processes the raw textual content, while the graph model captures the structure induced by the broader learning ecosystem. In this view, GEML can be seen as an early instantiation of graph-based learner modeling, where structured relations complement content-based signals. The mutual learning mechanism naturally enables the graph-centric and text-centric views of a learner to exchange information at the probability level, without forcing them into a single monolithic model.

A central theme of this thesis is multimodal learner modeling, where different modalities, such as text,

behavioural logs, sensor data, skeletal motion, and others, are combined with large pre-trained models to derive richer representations of learners. The GEML framework offers a reusable template for such integration:

- Each modality or view is modeled by a dedicated backbone (PLM, spatio-temporal GNN, or multimodal large language model);
- A graph captures structural relations among entities (learners, activities, resources), possibly across modalities;
- Mutual learning terms couple the predictive distributions of these backbones, enabling cross-modal knowledge transfer while keeping the architectures loosely coupled.

In later chapters, this idea is extended beyond text to multimodal data. Although the specific architectures differ, the conceptual pattern—separate channels, shared probability space, and uncertainty-aware coupling—directly echoes GEML.

While GEML demonstrates strong performance and conceptual elegance, it also has limitations that motivate subsequent work. The experiments are conducted on general NLP benchmarks rather than domain-specific educational datasets. Applying GEML to real-world learner data would require careful graph construction and potentially task-specific adaptations; The framework currently considers two learners. Extending mutual learning to multiple heterogeneous learners, such as text, sensor, and video backbones, raises questions about the scalability and stability of the coupling, which this thesis explores in multimodal settings. GEML assumes centralized training and does not address privacy constraints. Federated variants of graph-enhanced mutual learning, where graph and sequence models are distributed across clients, represent a natural but non-trivial extension. These limitations are not shortcomings of GEML, but rather indicators that it occupies a specific point in a broader design space. Subsequent chapters build on GEML's principles to develop multimodal and federated frameworks, such as skeleton-based action recognition under clustered federated learning, that bring the thesis closer to real-world educational applications while preserving the core idea of "loosely coupled, strongly synergistic" graph-enhanced mutual learning.

In summary, this chapter has presented GEML as both a concrete contribution to text classification and a conceptual bridge towards the more complex multimodal and federated systems that form the core of this work. By decoupling architectures while coupling probability distributions, GEML provides a reusable and extensible pattern for integrating graph structure with powerful pre-trained models, a pattern that will reappear in various forms throughout the remainder of the thesis.

5

CF-STGAT: Clustered Federated Spatio-Temporal Graph Attention Networks

This chapter introduces the CF-STGAT framework, which leverages clustered Federated learning (FL) to improve skeleton-based action recognition under strong client heterogeneity. On the one hand, CF-STGAT provides a unified, structure-aware federated baseline built on a shared STGAT backbone, systematically addressing the mismatch between non-IID client data and global aggregation that can lead to client drift and unstable convergence under standard FedAvg. On the other hand, it proposes a general cluster-then-regularize paradigm: (1) clients are dynamically grouped in a data-driven manner and (2) the resulting cluster models are lightly coordinated. The former uses compact, model-internal descriptors derived from attention statistics, and the latter is coordinated through an attention-similarity-weighted global reference and a regularized fusion step that aligns clusters without collapsing their specialization. In this way, CF-STGAT demonstrates how robust personalization and stable optimization can be achieved by combining structure-informed grouping with cross-group coordination, while preserving practical deployability by keeping local training unchanged and introducing only server-side statistics extraction, clustering, and lightweight fusion. This design philosophy of being modular yet strongly coordinated will be extended in later chapters to broader federated and graph-centric learning settings, including clustered federated learning for multimodal learner modeling and federated graph learning scenarios.

5.1 Motivation

Skeleton-based human action recognition (HAR) is a representative instance of learning where structure is not merely a helpful prior but the primary organizing principle of the signal. A skeleton sequence can be formalized as a time-indexed graph signal: joints constitute nodes, anatomical relations induce edges, and the classification objective requires capturing both pose configurations and their evolution over time. This viewpoint has driven the field away from purely sequence-centric encoders toward spatio-temporal graph models in which relational computation is explicitly embedded in the backbone. Among such models, attention-based designs are especially attractive because they can expand the effective receptive field while preserving topology-aware computation. In particular, Spatio-Temporal Graph Attention Network (STGAT) provides an interpretable mechanism that learns to allocate importance across joints and temporal neighborhoods, yielding a compact summary of learned dependency structure [90]. From a deployment standpoint, skeleton streams are rarely centralized. They typically originate from distributed devices such as RGB-D cameras, wearables, or distinct organizations like laboratories, hospitals, schools,

and privacy policies frequently prohibit collecting raw sequences on a central server. FL offers a practical collaboration interface under these constraints: raw data remain local while clients exchange model updates, enabling joint optimization without direct data sharing [91]. A direct baseline Federated STGAT (Fed-STGAT) in this setting integrates STGAT with FedAvg, providing a minimal pipeline that respects the privacy boundary and is easy to deploy. As shown in Fig. 6, each client trains a local STGAT on its private skeleton sequences and periodically uploads model updates. The server aggregates the received updates to obtain the global model and broadcasts it back to clients for the next round. Raw data never leaves the clients, preserving data privacy. The primary obstacle, however, is that the implicit assumptions behind naive averaging are systematically violated. Non-IID heterogeneity in skeleton HAR is structural rather than incidental: cross-subject variation changes motion style and kinematic patterns even within the same action class, and cross-setup variation changes viewpoint and capture geometry. When client objectives differ substantially, local updates can point toward different optima; averaging them can induce client drift, manifested as oscillatory convergence, unstable global generalization, and high variance across clients [91]. Regularization-based methods such as FedProx constrain local deviations from the broadcast model [92], and contrastive methods such as MOON attempt to align representations across clients [93]. Variance-reduction and normalization strategies further correct client-specific bias and objective inconsistency under heterogeneous local updates [94, 95]. These approaches can mitigate drift while preserving a single global model, but they do not change the fundamental aggregation topology, which can remain brittle under large heterogeneity.

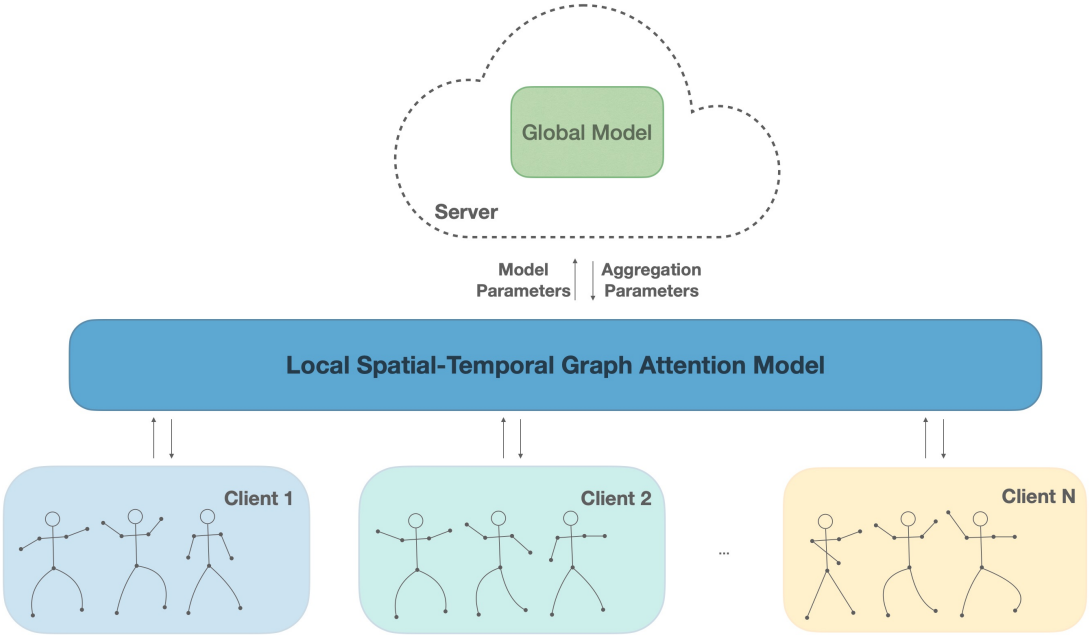


Figure 6: Overview of Fed-STGAT.

Clustered federated learning addresses the topology by grouping similar clients and training a model per group. Conceptually, this reduces within-group gradient variance and supports group-level personalization

[96]. In practice, its effectiveness depends on the quality of the similarity signal used for grouping. Many clustering strategies operate on a generic update geometry that may be weakly aligned with the domain structure that actually drives performance in graph-based skeleton HAR. Prior clustered baselines for HAR reflect this generic-similarity paradigm and motivate structure-aware alternatives [97]. CF-STGAT is motivated by a thesis-consistent premise: *when the learning problem is explicitly graph-structured, group formation should be driven by structure-aware signals, and the resulting groups should be coordinated to avoid excessive fragmentation.*

CF-STGAT operationalizes this premise by exploiting a model-internal signal that is both structure-aligned and communication-compatible. Multi-head spatio-temporal attention parameters encode how the model allocates importance across joints and temporal neighborhoods. Because these parameters are already part of the communicated model state, the server can extract lightweight attention-derived descriptors without requiring raw data, per-sample activations, or additional client-side objectives. CF-STGAT then forms dynamic clusters based on these descriptors, performs cluster-wise aggregation to reduce drift, and introduces a similarity-weighted fusion mechanism to regularize inter-cluster divergence. The resulting argument chain is coherent: attention-derived similarity improves grouping quality; clustering reduces within-group drift; fusion limits between-group drift; together they improve accuracy and stability under strong heterogeneity.

5.2 CF-STGAT Framework

We adopt STGAT as the local model due to its strong empirical performance in capturing short-term features for skeleton sequences via spatio-temporal patterns. Fig. 7 below summarizes the overall CF-STGAT pipeline and clarifies how local training, server-side clustering, and cross-cluster coordination are integrated without changing the client optimization procedure. Concretely, in each communication round, every client n performs E local update steps on its STGAT model and uploads the resulting parameters $\boldsymbol{\theta}_n$ to the server. On the server, attention statistics are extracted from the received models, normalized, and aligned across attention heads to obtain comparable client descriptors; every I_c rounds, these descriptors are used to dynamically partition clients into clusters $\{C_1, \ldots, C_C\}$ via reclustering. Given the current clustering, the server then performs inter-cluster aggregation to form an attention-similarity-weighted global reference model $\boldsymbol{w}_G$, and regularizes each cluster model through a lightweight fusion update, which gently aligns cluster models while preserving their specialization. The updated cluster models are broadcast back to clients for the next round, ensuring that personalization arises primarily from structure-aware grouping, while stability is promoted through controlled cross-cluster coordination.

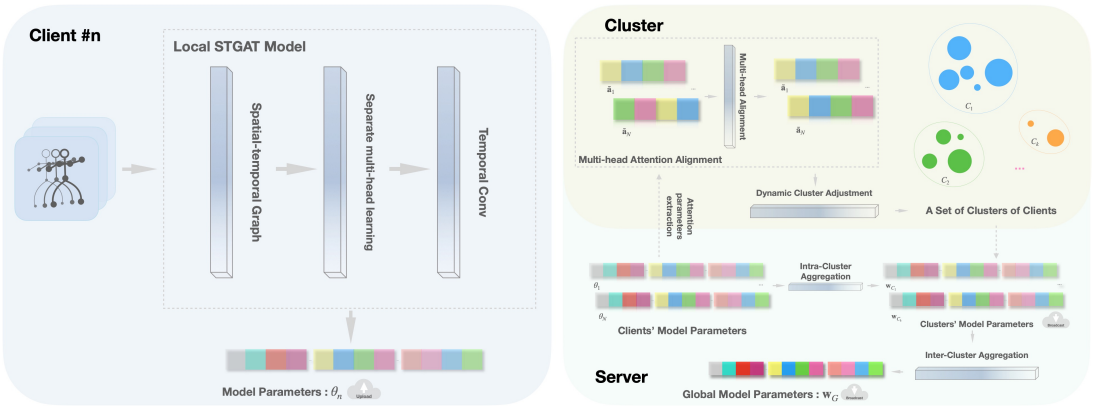


Figure 7: CF-STGAT pipeline.

5.2.1 Preliminaries

We adopt the standard spatio-temporal skeleton graph abstraction. A skeleton at each frame is a graph $G = (V, A)$ where V denotes joints and A encodes anatomical adjacency. A sequence is represented as a tensor $X \in \mathbb{R}^{C \times T \times N}$, where N is the number of joints, T the number of frames, and C the feature dimension per joint. Temporal context is incorporated by connecting joints across adjacent frames so that message passing can mix spatial relations and short-horizon temporal relations. STGAT instantiates this computation through spatio-temporal attention: for each timestamp, the model assembles a local spatio-temporal neighborhood and uses multi-head attention to weight neighbors adaptively, emphasizing informative joints and time offsets while attenuating less relevant connections. STGAT plays two roles in CF-STGAT. It is the local backbone trained on each client, and its attention parameters constitute the source of client descriptors for structure-aware clustering.

Fed-STGAT follows the canonical FL protocol. At round t, the server broadcasts the global model $\Theta^{(t)}$ to selected clients. Each client initializes its local model with $\Theta^{(t)}$, performs local optimization on its private dataset D_i for a limited budget, and uploads the updated parameters $\Theta_i^{(t+1)}$. The server aggregates updates using sample-size weights to form $\Theta^{(t+1)}$. This baseline is minimal and preserves the privacy boundary by design. Its limitation is structural: when client distributions are heterogeneous, the aggregated update can be inconsistent with many clients simultaneously, yielding oscillations, slow progress, and unstable generalization. In skeleton HAR, this effect is amplified because heterogeneity changes the dependency structure itself (how joints and time interact) rather than merely shifting marginal feature distributions.

5.2.2 Dynamic clustering and cluster-wise aggregation

CF-STGAT repurposes STGAT attention parameters as a structure-aware similarity signal. After clients upload their model parameters, the server extracts multi-head attention statistics from the attention layers

and concatenates them into a descriptor that summarizes how each client's training has shaped the model's spatio-temporal dependency preferences. The key conceptual shift is that clustering is performed on a domain-aligned summary of relational modeling behavior rather than on generic parameter deltas. To stabilize this feature space, CF-STGAT applies Z-score normalization to each descriptor and then performs PCA on the server, selecting the reduced dimension so that 90% of cumulative explained variance is retained. The resulting embedding z_i preserves the majority of informative variation while reducing noise and computational cost for clustering.

Given embeddings z_i, CF-STGAT periodically reclusters clients using K-means. Reclustering occurs every I_c communication rounds. The dynamic nature is essential: clients are not permanently assigned to a cluster; instead, cluster membership can evolve as local models and the decision boundary changes throughout training. The server thereby maintains multiple cluster models $\bar{w}_{C_k}$ rather than a single global model. After clusters are formed, aggregation proceeds within each cluster. For a cluster C_k, the server computes a size-weighted average of the uploaded client models to obtain $\bar{w}_{C_k}$. This cluster-wise aggregation can be interpreted as reducing within-group gradient variance. Because clients within a cluster are selected to have similar attention-derived dependency patterns, their updates are more aligned, and the cluster aggregate is less noisy than a global aggregate.

5.2.3 Similarity-weighted inter-cluster fusion

Clustering improves personalization but can also lead to fragmentation if cluster models drift too far apart. CF-STGAT addresses this risk through a lightweight fusion step that softly couples cluster models via a reference model guided by attention similarity. Let a_k denote the attention centroid for cluster C_k, and let a_G be a global attention prototype. The server computes cosine similarities $s_k = \cos(a_k, a_G)$ and defines weights that reflect both similarity and cluster size:

$$\alpha_k = \frac{N_k s_k}{\sum_j N_j s_j}, \qquad \sum_k \alpha_k = 1, \tag{5.1}$$

where N_k is the number of clients in C_k. A reference model is constructed by mixing cluster models:

$$w_G = \sum_k \alpha_k \bar{w} C_k. \tag{5.2}$$

Each cluster model is then softly updated toward this reference:

$$\bar{w} C_k \leftarrow (1 - \gamma)\bar{w}_{C_k} + \gamma w_G, \tag{5.3}$$

where $\gamma \in [0, 1]$ controls the fusion strength. The intended effect is not to collapse all clusters into a single model but to preserve specialization while limiting pathological inter-cluster divergence. The method further describes how to adapt γ based on model discrepancy, i.e., the fusion strengthens when divergence is larger and relaxes when alignment is already sufficient.

A practical merit of CF-STGAT is that it does not require modifying client-side training. Clients train STGAT as usual; the additional logic is primarily server-side and executed intermittently. Descriptor extraction is applied to already-communicated parameters, and PCA/K-means are performed every I_c rounds in a reduced space. Under typical settings where descriptor dimension and the number of clusters are modest, this overhead is small relative to local training and network communication, making the approach compatible with realistic FL servers.

5.3 Experimental Evaluation

5.3.1 Datasets

CF-STGAT is evaluated on NTU RGB+D 60 (NTU 60) and NTU RGB+D 120 (NTU 120) under cross-subject (X-Sub) and cross-setup (X-Setup) protocols. NTU 60 contains 56,880 samples across 60 actions with 25 joints per frame. NTU 120 extends the action set to 120 with additional subjects and setups. To stress-test heterogeneity under FL, the evaluation uses a federated-by-dataset construction in which each (subject, setup) pair defines a client. This induces naturally non-IID label and domain distributions across clients. For NTU 60, an X-Setup protocol analogous to the NTU 120 cross-setup evaluation is additionally considered to explicitly stress setup shift. The resulting partition statistics are summarized in Table 4. The increase in client count and heterogeneity when moving from NTU 60 to NTU 120 provides a natural axis for interpreting the method's behavior, since the benefits of structure-aware clustering and fusion should scale with heterogeneity.

Datasets	Protocol	Classes	Train Clients	Test Clients
NTU 60	X-Sub	60	20	20
NTU 60	X-Setup	60	8	9
NTU 120	X-Sub	120	53	53
NTU 120	X-Setup	120	16	16

Table 4: Statistics of datasets.

5.3.2 Implementation details and Overall Results

All experiments use STGAT with eight graph attention layers and eight heads per layer. Sequences are processed by sampling 150 frames and randomly cropping to 128 frames for both training and testing. Federated training uses $E = 1$ local epoch per round for $R = 200$ communication rounds. The initial learning rate is 0.1 and is decayed by a factor of 10 at rounds $60, 90, 120$. Clustering is performed every $I_c = 5$ rounds, with $C = 4$ clusters on NTU 60 and $C = 6$ clusters on NTU 120. The fusion coefficient is selected from $\gamma \in 0.1, 0.2, 0.5$ via a small grid search.

Comparative baselines include the FedAvg-based Fed-STGAT, FedProx, MOON, and the clustered baseline FedCLAR. This set contrasts optimizer-level heterogeneity mitigation with topology-level grouping, and

isolates the incremental contribution of attention-guided clustering plus similarity-weighted fusion. Table 5 reports top-1 accuracy. CF-STGAT achieves the best performance in all settings. Relative to Fed-STGAT, the absolute gains are +0.84 on NTU 60 (X-Sub), +4.09 on NTU 60 (X-Setup), +7.98 on NTU 120 (X-Sub), and +4.18 on NTU 120 (X-Setup). The largest gain occurs on NTU 120 (X-Sub), consistent with the expectation that as heterogeneity becomes more severe, structure-aware grouping and controlled cross-group transfer provide increasing benefit. The method also improves over FedCLAR across settings, supporting the view that fusion contributes beyond clustering alone by constraining inter-cluster drift.

Model (STGAT backbone)	**NTU60 X-Sub**	**NTU60 X-Setup**	**NTU120 X-Sub**	**NTU120 X-Setup**
Fed-STGAT (FedAvg)	89.72	86.46	72.35	78.71
FedProx	89.31 (−0.41)	87.49 (+1.03)	75.36 (+3.01)	80.65 (+1.94)
MOON	88.01 (−1.71)	86.96 (+0.50)	73.31 (+1.04)	79.86 (+1.15)
FedCLAR	89.88 (+0.16)	87.24 (+0.78)	75.55 (+3.20)	80.83 (+2.12)
CF-STGAT	**90.56 (+0.84)**	**90.55 (+4.09)**	**80.33 (+7.98)**	**82.89 (+4.18)**

Table 5: Overall comparison (Top-1 accuracy, %).

5.3.3 Clustering stability

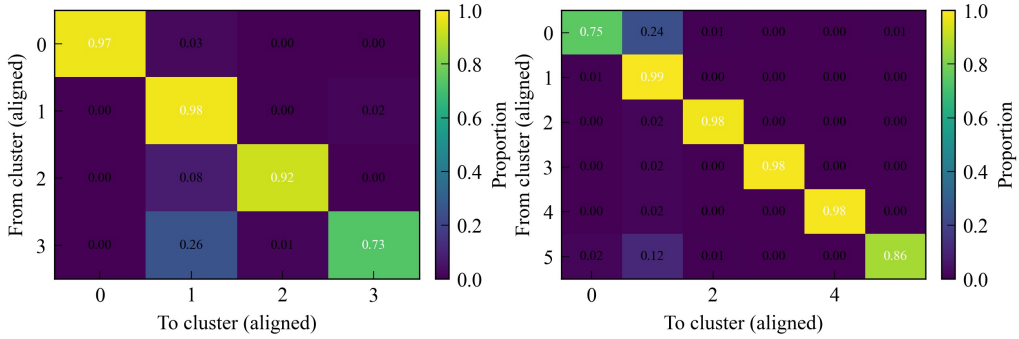
Dynamic clustering is only useful if it remains stable rather than oscillatory. CF-STGAT reports the Adjusted Rand Index (ARI) between consecutive reclustering checkpoints and the churn rate, which measures the fraction of clients changing clusters. The NTU 60 (X-Sub) shows in Table 6, the configuration $I_c = 5$ and $C = 4$ yields the strongest agreement (ARI $= 0.846 \pm 0.213$) and the lowest churn (0.040 ± 0.056). The accompanying interpretation is that too few clusters can conflate heterogeneous clients, while overly frequent reclustering can introduce transient boundary oscillations that depress ARI. The stability evidence therefore supports treating (I_c, C) as controls that influence not only accuracy but also the reliability of the group-formation process.

Setup	**ARI(mean ± std)**	**Churn(mean ± std)**
$I_c = 5, C = 4$	0.846 ± 0.213	0.040 ± 0.056
$I_c = 5, C = 3$	0.674 ± 0.326	0.064 ± 0.077
$I_c = 3, C = 4$	0.750 ± 0.236	0.064 ± 0.064

Table 6: Round-to-round clustering stability on NTU 60 (X-Sub)

In addition, we visualize the row-normalized average transition matrix to reveal where changes concentrate. The transition matrices in Figure 8 below are strongly diagonal on both datasets, confirming persistent cluster identities. On NTU 120, clusters 1–4 exhibit near-perfect self-transition (≥0.98), with localized drift from $0 \longrightarrow 1$ (0.24) and $5 \longrightarrow 1$ (0.12), that means a single fuzzy boundary while the rest remain stable. On NTU 60, overall stability is also high, but the boundary between clusters shows more activity: $3 \longrightarrow 1 \approx 0.26$ and $2 \longrightarrow 1 \approx 0.08$, consistent with the slightly lower ARI mean and the small but non-zero churn pulses. Overall, the table and heatmaps show that instability is concentrated at a few

adjacent-cluster boundaries rather than reflecting global oscillations.



(a) Average Transition | NTU 60 (X-Sub) (b) Average Transition | NTU 120 (X-Sub)

Figure 8: Row-normalized average transition matrices across consecutive checkpoints.

5.3.4 Mechanistic Interpretation

As shown in Figure 9 below, in the absence of clustering (treating all clients as a single group), the baseline method FedAvg and its regularized extensions FedProx and MOON exhibit slow and uneven convergence under non-IID partitions. Training a single global model with FedAvg on highly non-IID clients can induce large variance in the aggregated update, since local gradients are optimized for very different data distributions. This variance manifests as oscillatory behavior of the global model and unstable performance across communication rounds. Although MOON accelerates the initial rise, the jagged behavior remains, indicating that optimizer-level regularization alone is insufficient to eliminate client drift.

Consistently, the test loss in Fig. 10 below decreases more slowly and stabilizes at a higher asymptote under FedAvg-style training, with the effect particularly pronounced on NTU 120, where label and domain shifts are stronger. This behavior is consistent with uneven per-client convergence under non-IID partitions: without grouping, FedAvg and its regularized extensions aggregate high-variance local updates, which induce client drift and produce noisy, misaligned training trajectories. By contrast, our attention-guided clustering groups clients with similar attention-based descriptors, so that each cluster-wise model is optimized on a more homogeneous subset of clients; within a cluster, local gradients are better aligned, the variance of the aggregated update is reduced, and per-client trajectories become tighter, yielding smoother and more stable convergence than a single global model. Notably, clustering alone can still leave a persistent low-accuracy tail, which is indicative of residual inter-cluster divergence. The similarity-weighted fusion step therefore acts as a stabilizer: by constructing an attention-informed reference model and gently aligning each cluster model toward it, fusion tends to compress the long tail, raise the performance plateau, and reduce terminal test loss, with the clearest separation again observed in the more heterogeneous NTU 120 setting. Overall, the empirical trajectories support the intended complementarity: clustering primarily mitigates within-group drift, while fusion primarily limits between-group drift.

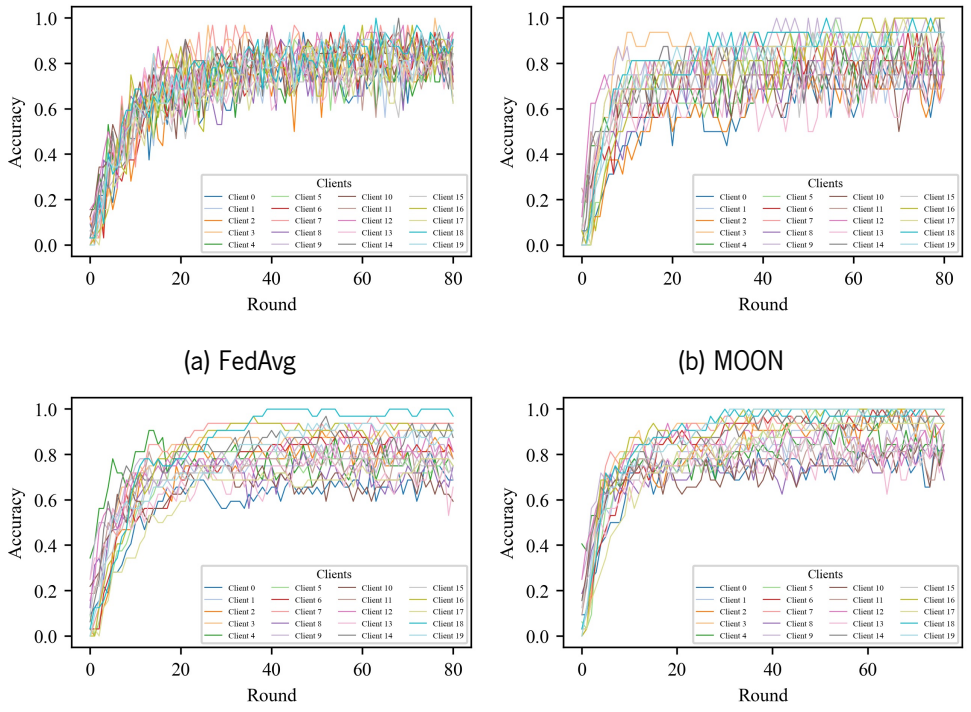


Figure 9: Per-client training accuracy vs. communication rounds on NTU RGB+D 60 (X-Sub).

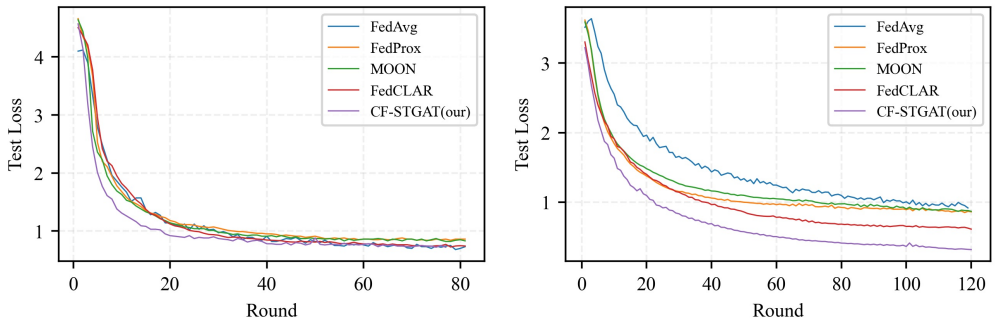


Figure 10: Test loss vs. communication rounds.

5.4 Discussion and Connection to the Thesis

The thesis theme is that group formation in decentralized learning should be derived from data/model behavior rather than imposed by metadata or assumed similarity. CF-STGAT provides a concrete instance of this principle by converting attention parameters into a client descriptor that encodes relational modeling behavior. It also contributes a coordination primitive that is reusable beyond skeleton HAR. The similarity-weighted reference model and soft fusion mitigate a fundamental tension in clustered FL: without coupling, specialization can become fragmentation; with overly strong coupling, personalization

collapses back to a single global model. By guiding coupling through attention similarity, CF-STGAT provides an interpretable mechanism that can be generalized to other graph-based learner modeling settings where attention-like modules encode how information flows across entities and time. The method is naturally composable with client-level personalization strategies. CF-STGAT personalizes at the cluster level by discovering shared structure among clients, whereas approaches such as FedPer/FedRep personalize at the client level by decoupling shared and private components. A thesis-level roadmap can therefore motivate hierarchical personalization in which cluster-level sharing provides a meso-scale inductive bias and client-specific heads or adapters provide micro-scale calibration within each cluster. This composition also motivates treating C, I_c, and γ as system-level control variables that jointly influence accuracy, stability, fairness (through long-tail compression), and cost, enabling a multi-objective design narrative aligned with the dissertation's system-and-algorithm perspective.

Using attention parameters as clustering features, CF-STGAT supports the integration of MLLMs and GNNs for grouping by showing that attention statistics from relational modules can provide compact, semantically meaningful, structure-aware descriptors for cohort discovery. In the thesis setting, where multimodal data are first encoded and then organized by a GNN to capture learner–content–concept relations, the resulting edge/node attentions provide a principled basis for grouping because they reflect how the model allocates importance across modalities, relations, and temporal dependencies, rather than relying on superficial metadata or raw feature distances. This establishes a concrete methodological link: *"grouping by model-internal relational signals" becomes a reusable design axis for MLLMs and GNNs grouping.*

6

Experiments for STEAM PBL Classroom Data Collection

This chapter presents a STEAM PBL classroom study on multimodal student data collection for intelligent grouping. The study was conducted at a Chinese elementary school from April to July 2022. The participants consisted of four STEM teachers, twenty grade-4 students, and the primary author of this study. In this setting, the study continuously collects evidence from both students' learning processes and learning outcome assessments, and operationalizes learner differences, collaboration fit, and grouping risks in ways that are observable and computable. The central objective of this chapter is to refine the workflow from classroom implementation to multi-source data collection and grouping outcome evaluation into a deployable experimental paradigm for intelligent grouping. On the one hand, it specifies what data should be collected by students and teachers, and how these data should be organized and annotated. On the other hand, it defines the control group setting, the experimental procedure, and the reportable forms of outcomes, ensuring that grouping-related features do not remain purely theoretical. Instead, they form a continuous evidence chain throughout project progression and further support dynamic grouping decisions and interpretable grouping optimization.

To achieve these goals, the chapter is organized into four layers: preparation, procedure, data, and evaluation. In the data preparation section 6.1, we describe the division of roles among participants, the control conditions, and the setup of the data collection infrastructure, including audio and video recordings, logs, artifacts, and assessments. At the same time, we detail the implementation rhythm of the four PBL projects, including phase milestones and key sampling points. In the experimental data section 6.2, we curate, align, and annotate multi-source data, including pre-study student questionnaires, in-process video and audio, project presentations and exhibit materials, outcome data, and teacher ratings, thereby forming a dataset that can be directly used for modeling and evaluation. And we present the evaluation metrics and comparison methods for grouping outcomes, specifying how to test the gains of intelligent grouping relative to baseline grouping while controlling for project-level differences, and how to produce results in forms suitable for reporting in this work.

6.1 Experimental Preparation

This section details the preparation work required to carry out an evidence-driven intelligent grouping experiment in an authentic primary-school STEAM PBL setting. Specifically, we establish a structure for

data collection and quality control, the control treatment design and implementation constraints, and a standardized multimodal data acquisition infrastructure (audio-video, artifacts, logs, questionnaires, and rubric-based assessments). The overarching principle is low intrusion. Data collection is embedded into routine classroom activities and relies on fixed camera placement and lightweight checklists, while fine-grained annotation and agreement checks are conducted after class. In doing so, the experiment ensures that multimodal evidence remains continuously traceable, aligned across sources via anonymous IDs, and reusable for subsequent modeling and evaluation.

Teacher	(1) Deliver project-based instruction and organize classroom activities, ensuring consistency across projects in milestones and expected outputs. (2) Provide formative feedback and comments during presentations, offering process-oriented assessment cues and guiding group improvement. (3) End-of-class checklist logging: confirm within-group role allocation, and record conflicts, bottlenecks, and teacher interventions (yes/no, plus brief notes).
Observer / RA	(1) Conduct pre-class equipment checks and camera placement, ensuring audio/video capture quality and low-intrusion classroom operation. (2) Record key classroom time points: start/end time, phase milestone completion time, and timestamps of teacher intervention events; mark and index critical events (conflict, bottlenecks, help-seeking, and key iterations) with timestamps and brief notes. (3) Collect, scan, and archive post-class materials, ensuring cross-source alignment and traceable records.
Researcher	(1) Define the data dictionary, naming convention, and directory structure, standardizing data granularity, alignment keys, and missing-data handling rules. (2) Develop the annotation guideline and coding ontology, organize annotator training, and conduct sampled agreement checks and quality control. (3) Coordinate rubric scoring calibration, control scoring drift across projects, and maintain evaluation comparability. (4) Run the intelligent grouping workflow at project transition points, and produce interpretable grouping plans for teacher review and classroom deployment.

Table 7: Roles and responsibilities.

6.1.1 Roles and Responsibilities

To ensure sustainable multimodal data collection in authentic classroom settings, traceable archiving, and reproducible annotation, this study adopts a three-role division of labor, namely instructional delivery, collection and logging, and annotation quality control. The implementation follows a low intrusion principle, which embeds data capture and logging tasks into existing classroom routines as much as possible, thereby minimizing additional burden on student collaboration and project progression. The responsibilities and deliverables for each role are summarized in Table 7.

To ensure the stable production of alignable process evidence in primary school STEAM PBL, and to reduce noise introduced by unclear responsibility allocation in collaborative tasks, this study adopts a role-based design for within-group collaboration in each project. Specifically, each group is assigned core roles, including a leader, an engineer, a recorder, and a presenter. The implementation follows a "fixed within project, rotated across projects" strategy. Roles remain relatively stable within a project to preserve task continuity and evidence traceability, while rotating roles across projects helps prevent the solidification of ability labels and broadens the observation window for individual collaboration tendencies. Table 8 summarizes the core responsibilities of each role, providing an operational basis for subsequent process data collection and feature construction for intelligent grouping.

Leader	(1) Organize in-group discussions and decision-making, and drive progress toward phase milestones. (2) Coordinate task allocation and resources, manage disagreements, and facilitate conflict repair. (3) Consolidate issues and requirements and communicate with the teacher, ensuring timely task completion.
Engineer	(1) Lead construction, electronic wiring, and programming implementation. (2) Drive debugging and test iterations, diagnose faults, and propose improvement plans. (3) Ensure functional stability and engineering feasibility of the artifact.
Recorder	(1) Complete and organize worksheets, engineering logs, and test records. (2) Document key rationales for decisions (requirements, trade-offs, failure causes, and improvement points). (3) Archive materials after class, ensuring process evidence is complete and traceable.
Presenter	(1) Lead presentation and Q&A, and organize presentation materials. (2) Clearly communicate the logic chain of "requirements-solution-testing-revision". (3) Consolidate peer and teacher feedback and drive group revision and reflection.

Table 8: Member roles and responsibilities in a group.

6.1.2 Four STEAM PBL Projects

This study implemented four Maker projects across four rounds of Research Lessons. Under the DTMP framework, students progressed through a full project-based learning cycle of "understanding contextual needs, ideation, prototyping, test-driven iteration, and presentation and reflection."The four projects were situated in authentic scenarios closely connected to students' lives, including campus services, learning habits, and parent-child communication. Under the same classroom organization and evidence collection mechanisms (e.g., worksheets, test logs, presentations, and reflections), each project continuously

produced alignable process evidence, providing a consistent data basis for subsequent comparisons of grouping strategies:

- **Project 1 (P1): Smart Elevator Button** This project is situated in a campus service scenario and asks students to design a smart elevator button for a real need, namely facilitating elevator use for cafeteria staff. Students collect and summarize user needs through interviews, generate design ideas under material constraints, implement core functions by connecting basic components such as a button and digital display, document trial and error and improvements during prototyping, and finalize the project with reflection and summary.
- **Project 2 (P2): Smart Bags** This project focuses on students' everyday use contexts and requires the design of a multifunctional smart backpack, such as correcting hunched posture, using a light sensor to remind students to close the zipper, and optionally exploring additional functions. The project emphasizes understanding and describing sensor functions, proposing solutions under constraints, completing wiring and programming, documenting testing, failures, and refinements, and reflecting on the Design Thinking process.
- **Project 3 (P3): Smart Phone Holder** This project targets time management in smartphone use and requires a clear interaction logic: picking up the phone triggers timing, exceeding the time limit triggers a buzzer alarm, and placing the phone back stops the alarm. Building on prior practice with button wiring, programming, and testing, students complete the program design. During ideation, they consider material and capability constraints, evaluate and refine sketches and form factors, and conclude with a summary of the Design Thinking process.
- **Project 4 (P4): Smart Parent-Child Communicator** This project is situated in a parent-child communication scenario and asks students to design a smart communication device. Based on buttons and LEDs, students implement information encoding and expression, such as using a single LED to convey a message, using two LEDs to represent multiple messages, and optionally exploring additional functions. The project includes revisiting basic wiring and testing, identifying root causes of communication problems, proposing solutions under constraints, documenting and explaining prototype issues and their causes, completing sketch designs that consider size, aesthetics, and usability, and concluding with a reflection on both project value and the Design Thinking process.

Importantly, the four projects were designed as a unified project carrier for controlled comparisons. The goal was not to create a sequence with systematically increasing difficulty, but to keep all projects within the same difficulty level and the same range of required competencies. All four projects are bound to basic interaction and control tasks that are feasible for Grade-4 students. They follow the shared cycle of "ideation, prototyping, test-driven iteration, and presentation and reflection". Also, they are implemented with the same class time structure, unified materials and tool constraints, and a consistent rubric-based evaluation framework. Therefore, differences across projects mainly arise from the problem context and

functional narrative, rather than from systematic differences in knowledge depth or engineering complexity. This design enables the study to treat grouping strategy (random, teacher-based, intelligent, and iterative intelligent grouping) as the primary independent variable, and to control the influence of project differences through consistent data collection and evaluation protocols across projects.

6.1.3 Control and Treatments

To ensure that differences in grouping strategy serve as the primary independent variable, this study adopts a within-class sequential quasi-experiment design, organizing four STEAM projects in a structure "two baseline controls followed by two treatments". The core rationale is to accumulate sufficient process evidence from the first two projects in authentic classroom settings to support subsequent modeling and inference for intelligent grouping. Then, without changing the instructional content framework or the evaluation system, the grouping strategy is introduced as the only systematic source of variation, enabling interpretable and reproducible controlled validation. The grouping conditions are as follows:

- **Control A: Random Grouping** Before the project begins, the research assistant assigns students randomly based on anonymous IDs. This condition provides baseline data under natural collaboration variability, yielding the process evidence and outcome distribution observed under no prior intervention.
- **Control B: Teacher Experience-Based Grouping** The teacher forms groups based on prior classroom observations and academic performance, potentially considering ability balancing and behavior management. This condition characterizes a common practice in real classrooms and provides a second source of authentic grouping evidence for subsequent intelligent grouping.
- **Treatment 1: Intelligent Grouping** Using the accumulated data from P1 and P2, the study constructs student profiles and a learner graph to generate candidate grouping plans (at least two alternatives, A and B) with interpretable rationales, such as ability balancing, role complementarity, and conflict risk minimization. The final grouping is determined and implemented after a teacher-in-the-loop review during lesson preparation.
- **Treatment 2: Iterative Intelligent Grouping** Building on P1 and P2, the study further integrates P3 data to update student profiles and the learner graph, and performs rolling optimization of grouping plans at the project transition point. This condition tests whether profile updating and model iteration yield incremental improvement.

To reduce confounding effects from project themes, teacher differences, or resource variation, the following key conditions are constrained throughout the study:

All four projects follow the same PBL workflow and engineering cycles: design, build, test, refine, and

present, and use consistent phase milestones and deliverable formats (worksheets, engineering logs, artifacts, and presentations), thereby limiting project-level differences in the comparison. A single rubric framework is used throughout to score both product quality and process quality. After each project, a structurally consistent learning outcome assessment is administered, ensuring comparable outcome indicators across conditions. All groups use the same material kits and equipment support. Teacher intervention follows a unified facilitation principle, and intervention frequency and types are logged as process control variables and explanatory variables. Each group adopts a unified set of student roles (leader, engineer, recorder, and presenter) that remain stable within each project and rotate across projects. This reduces collaboration noise caused by ambiguous responsibility and provides a stable observation window for grouping related features.

This Control–Treatment organization serves two goals simultaneously. First, it provides comparison baselines across three representative grouping types: random, teacher experience-based, and intelligent grouping, enabling tests of the relative advantages of intelligent grouping. Second, it forms a closed loop of "data accumulation, profile modeling, grouping decision, and re-collection with iteration," so that grouping strategies can be continuously updated with process evidence in authentic classrooms, thereby supporting empirical validation of dynamic grouping and targeted instructional support.

6.1.4 Data Infrastructure and Questionnaires

This study builds a multimodal evidence chain-oriented data collection infrastructure to satisfy two requirements simultaneously: feasibility in authentic classrooms and computability for intelligent grouping. The design follows three core principles: low intrusion, alignability, and traceability. Low intrusion means that data capture is embedded into routine classroom workflows. Alignability requires a shared index and timeline across heterogeneous data sources. Traceability ensures that every label and derived feature can be traced back to its original evidence location in the raw data.

To ensure consistent alignment and traceability across heterogeneous data sources, including student questionnaires, classroom audio-video, and teacher evaluation records, this study adopts a unified indexing scheme based on three core identifiers: student ID (`sid`), project ID (`pid`), and group ID (`gid`). The `sid` serves as an anonymized individual-level key that links dispositional priors, student-level features, and individual contribution assessments. The `pid` identifies the instructional project (P1–P4) and encodes the experimental phase and grouping condition, thereby enabling project-level control in subsequent analyses. The `gid` captures collaborative structure by associating students with their working groups and supports the aggregation of process features, group-level rubric scores, and outcomes. Together, these identifiers form a global primary key system that allows all structured tables to be joined and queried consistently, and enables model inputs and outputs to be traced back to the corresponding classroom evidence. Table 9 below summarizes the key index fields used throughout the dataset.

In the experimental preparation phase, the classroom intervention was structured around the Design

Field	Type	Example	Description
sid	str	S01-S20	Student ID.
pid	str	P1-P4	Project ID.
gid	str	G1-G5	Group ID.

Table 9: Key index.

Thinking (DT) framework proposed by the Hasso Plattner Institute of Design at Stanford (d.school), which organises project-based learning into five stages: Empathize, Define, Ideate, Prototype, and Test. Importantly, this stage-based structure provides a pedagogically meaningful backbone for intelligent grouping, because it yields comparable and time-anchored evidence of students' collaboration and problem-solving processes across projects. Specifically, Empathize and Define externalise students' abilities to elicit user needs and articulate problem statements through structured inquiry 5W+H(Who, What, When, Why, Where, and How), which can be operationalised as early indicators of communication quality, perspective-taking, and analytical reasoning. Ideate activities, supported by tools such as a 2×2 Matrix, make group negotiation, idea diversity, and prioritisation strategies observable, enabling the extraction of interaction patterns relevant to role allocation and compatibility. The Prototype and Test phases, where sketches and mock-ups are iteratively produced and evaluated in context, generate dense process evidence of persistence, coordination, iterative debugging, and feedback utilization. Therefore, the DT stages and their associated activities were not only used to guide instruction, but also to standardise data collection and indexing, so that student profiles and peer-relation signals for intelligent grouping could be derived from aligned, stage-consistent behavioural evidence rather than from final products alone:

- **Audio and Video Capture** Audio and video data capture collaborative interaction and problem-solving processes, and serve as a key source for constructing features related to collaboration fit and grouping risk. Considering equipment feasibility and coverage needs in primary school classrooms, we adopt a minimum viable configuration of group views and classroom views. *Group close-up views:* Each group is recorded by a fixed wide-angle camera that covers discussions and making activities, preferably including both tabletop materials and partial facial profiles, so that role enactment, negotiation, hands-on implementation, and help-seeking behaviors can be observed. *Classroom overview:* The full classroom view records classroom order, teacher circulation trajectories, cross-group interaction, and teacher intervention events, enabling aligned analysis of teacher support intensity and intervention types. To strengthen alignment, each session begins with a shared synchronization marker as shown in Table 10 and Table 11 below. In addition, timestamps are recorded at key phase transition points, forming coarse-grained time anchors for subsequent indexing.

- **Structured Classroom Logs** Structured logs translate observable events into computable fields, reducing reliance on frame-level video annotation and providing low-cost features for modeling.

Logs are implemented as checkbox-based forms for rapid recording and include at least four categories of fields. *Timeline fields:* start and end times, phase transition times, and milestone completion times, for example, solution finalized, prototype completed, first test conducted, first iteration completed, and final presentation prepared. *Role allocation fields:* the primary role holders for the session, whether role allocation changed, and whether role vacancy or task stacking occurred. *Teacher intervention fields:* frequency and types of interventions (prompting, guiding questions, demonstration, and direct correction), and triggers (technical bottlenecks, collaboration conflicts, and time management issues). *Critical event fields:* conflict events, bottlenecks, help-seeking events, and key iteration events, each recorded using a "yes or no, brief note, and timestamp" format. These logs serve both as direct sources for process indicators (e.g., intervention frequency, iteration count, and conflict rate) and as indices for efficient video review, enabling a practical balance between executability and information density.

Field	Type	Example	Description
pid	str	P3	Project ID.
uid	str	P3_U01234	Utterance ID.
sid	str	S07	Speaker (student ID).
gid	str	G2	Group at the time (based on the grouping plan).
t_start	float	523.4	Start time in seconds.
t_end	float	531.8	End time in seconds.
duration	float	8.4	Duration in seconds.
text	str	...	Transcribed text.

Table 10: Audio transcript Utterance-Level schema.

Field	Type	Example	Description
pid	str	P3	Project ID.
src_sid	str	S07	Source student (initiator).
dst_sid	str	S03	Target student (replied-to/mentioned/helped).
edge_type	str	reply/mention/help	Edge type.
w_count	int	6	Interaction count.
w_time	float	35.2	Total duration of interaction-related utterances.

Table 11: Interaction edge schema.

Before the PBL class, each student did the pre-study questionnaire as follows:

- **Artifacts and Process Products** Artifacts document what students built, how they built it, and the achieved quality level, and thus provide the core evidence linking process to outcomes. To ensure cross-project comparability and cross-source alignment, artifacts follow a unified archiving

structure and naming convention, indexed along four levels: project, group, phase, and individual. *Design and ideation materials:* interview notes and needs summaries, design sketches, and records of trade-off rationales. *Implementation and testing materials:* engineering logs, wiring and code records (photos or screenshots), and test logs including test time, description, pass or fail, failure causes, and improvement notes. *Outcomes and communication materials:* photos or videos of final products, presentation materials (posters, slides, and scripts), and optional records of Q&A questions and response highlights. All paper-based materials are scanned and archived after class. Digital materials (photos and screenshots) are backed up and verified on the same day to prevent data loss due to device changes or accidental deletion.

- **Learning Outcome Assessments** Learning outcome data provide quantitative references for ability complementarity and learning gains, and serve as primary labels for evaluating the effectiveness of intelligent grouping. Four assessment signals are used consistently across projects. *Project tests:* Each project concludes with a post-test, and if needed, a very short pre-test. Tests cover core knowledge and key skills. To control for differences in item difficulty across projects, test scores are standardized in later analyses. *Rubric-based scoring:* A unified rubric is used to evaluate both product quality and process quality, producing group-level scores and optional individual contribution scores. Rater calibration and sampled re-scoring are used to control drift. *Questionnaires:* Core competency and collaboration tendency questionnaires are administered before and after the study, and when needed, short end-of-project forms are used, providing cold start features for grouping and outcome variables that may transfer across projects. *Peer assessment:* A brief end-of-project peer rating captures contribution, communication, and collaboration satisfaction, supporting edge weight signals for the learner graph and assisting conflict risk detection.

- **Alignment and Data Governance** To ensure cross-source alignment and traceability, we adopt a three-layer alignment mechanism based on anonymous IDs, timestamps, and standardized directory structure and naming. *Anonymous ID alignment:* Student anonymous IDs and group IDs serve as primary keys to align questionnaires, tests, rubric scores, and artifacts. *Timestamp alignment:* Milestone timestamps from structured logs index key events and align them with audio and video segments, enabling a closed loop in which a table field can be traced back to the corresponding video evidence. *Directory structure and naming conventions:* Data are archived hierarchically under /P{k}/video/, /P{k}/logs/, /P{k}/artifacts/, and /P{k}/scores/. File names include project ID, group ID, student ID (for individual materials), and data type suffixes, enabling efficient retrieval and reproducibility. With this instrumentation, experimental outputs are not merely a collection of raw multimedia files. Instead, they constitute a structured evidence chain dataset that can directly support subsequent stages of feature extraction, grouping inference, and effectiveness evaluation.

6.2 Experimental Data

This section organizes multi-source classroom data into a modeling-ready structure of samples, labels, and metadata, and specifies unified rules for naming, alignment, and annotation. The goal of this specification is not to maximize collection complexity, but to establish a practical data closed loop under the constraints of authentic primary school classrooms. After completing Data Organization, Alignment, and executable Annotation, the classroom evidence chain is further packaged into a standardized dataset package that can be directly used for modeling and evaluation. The packaging process aims to convert heterogeneous evidence generated during classroom implementation into a unified data interface, enabling intelligent grouping algorithms to operate on a consistent input structure and allowing model outputs to be traced back to the original evidence for interpretable review. Overall, the workflow consists of four consecutive stages: raw data archiving, metadata indexing, data annotation, and standardized packaging. All four stages utilize student anonymous IDs, project IDs, and group IDs as unified alignment keys and employ event timestamps as location anchors for process evidence, thereby forming a traceable, alignable, and reusable closed-loop evidence chain dataset.

Dimension	**Item Content (Maker Mindset Pre-Study)**
Thinking Skills (TS)	1. I can test and compare different solutions and list their advantages and disadvantages. 2. I can summarize and organize the group's test data using charts or tables so that the results are clear. 3. I believe I can understand every knowledge point involved in Maker projects. 4. I believe I can master the knowledge involved in Maker activities. 5. I can propose a question that is worth investigating. 6. I believe that what I learn in Maker class can be applied to other classes. 7. I can use charts or tables to summarize and synthesize the group's test data. 8. I can design detailed steps to solve a problem.
Self-Efficacy (SE)	1. Compared with other classmates, I believe I can make more refined artifacts. 2. In Maker class, I believe I can complete more complex tasks. 3. Even after class, I still want to stay in the classroom to finish my project. 4. I like challenging class assignments because I can learn new things. 5. I believe I can complete the tasks assigned in class.
Learning Motivation (LM)	1. When I encounter problems during learning, I can think proactively rather than waiting for others to give answers. 2. No matter what subject I learn, I want to know more related content. 3. In the Maker class, compared with the final product, I prefer the process of making the product. 4. I think what I learn in Maker class is useful to me. 5. I think solving difficult problems is meaningful. 6. I think what we learn in Maker class is interesting.

Table 12: Pre-Study Questionnaire (Part I)

Dimension	Item Content (Maker Mindset Pre-Study)
Self-Concept (SC)	1. I think my work is highly innovative. 2. When working in a group, I think I can lead other members to achieve the goal. 3. When the task assigned by the teacher is very difficult, I will give up. 4. Even if my opinion differs from most people, I think I still dare to express my ideas. 5. When I make decisions, I think I will consider problems from others' perspectives.
Self-Regulation (SR)	1. When studying alone, I can use my time reasonably. 2. Compared with help from teachers or classmates, I prefer to complete projects independently. 3. When I encounter difficulties in a project, I can solve problems by searching for information or seeking help from the teacher. 4. When studying, I set goals for myself and monitor my progress. (Supplementary item) During self-study, I can use my time well. 5. When things do not develop the way I want, I can adjust my behavior to improve it.
Collaborative Learning (CL)	1. When I disagree with my group, I try to resolve the disagreement. 2. Whether in class or after class, I am willing to help other classmates so that they can do better. 3. I can share the results of my investigation with others. 4. If others' viewpoints are reasonable, I can accept them. 5. During group discussions, I can learn from others' strengths and supplement and correct my own viewpoints. 6. When I express my viewpoints, I explain my reasons.

Table 13: Pre-Study Questionnaire (Part II)

6.2.1 Raw Data Archiving

In this study, raw data archiving is not a simple storage operation. Instead, it is an evidence preservation process oriented toward educational empirical research and learning analytics. The goal is to preserve the temporal continuity, contextual integrity, and auditability of classroom evidence to the greatest extent possible, and to avoid introducing selection bias and interpretive drift at early processing stages. Based on the four-project classroom study workflow described above, the raw data consist of three core evidence sources: classroom audio-video process evidence, student questionnaire evidence as dispositional priors, and teacher evaluation evidence as expert judgment outcomes.

Audio-video evidence captures authentic group collaboration, hands-on making, teacher-student interactions, and key debugging behaviors. To maintain epistemic integrity, the archiving stage does not perform editing, segmentation, or pre-screening that retains only selected highlights. Instead, the original continuous timeline is preserved, and classroom phase anchors are retained as references for subsequent indexing and annotation. For each STEAM PBL project, 160-minute audio-video recordings were collected.

Questionnaire evidence comes from the Maker Mindset pre-study scale in Table 12 and Table 13 above. A total of 20 valid responses were collected, including 36 Likert items and 2 open-ended questions (37 items in total). Archiving uses item-level responses as the minimum granularity and retains raw responses before reverse scoring, enabling later scale validation, reverse coding, missing data handling, and multi-version aggregation, so that statistical and modeling procedures remain verifiable. Teacher evaluation evidence uses classroom observation and rubric-based descriptive rating standards as expert judgment evidence. Table 14 includes 7 evaluation stages. Six process and competency stages (STEM knowledge and skills, empathize, define, ideate, prototype, and share) use a five-level scheme (4/3/2/1/0), and the overall product score uses a four-level scheme (3/2/1/0). Archiving retains the descriptive text associated with each level, so that downstream label construction is not based only on numeric scores, but can also be traced back to the performance descriptions that justify each score, improving interpretability.

Stage	**Scoring Criteria**
STEM Knowledge and Skills	**4**: Students explore the functions of sensors, actuators, controller boards, and other basic components through experimentation, and create simple Maker projects. **3**: Students can distinguish the functions and uses of basic electronic components, and correctly use sensors, actuators, controller boards, and other basic components (matching, categorization, error analysis, summarization, and specific explanation). **2**: Students can identify, interpret, and explain the functions and uses of sensors, actuators, and controller boards, and can describe the logical relationships among them and other basic components. **1**: With teacher support, students accomplish part of the objectives. **0**: With teacher support, students do not accomplish the objectives.
Empathize	**4**: Students evaluate the necessity of the problem from the user's perspective based on collected data, justify its value, and clearly define the problem to be solved. **3**: Students summarize and categorize collected user needs, and list potential problems; they can identify and collect user needs via interviews, field observation, and iPad-based online search. **2**: Students can explain the concept of user needs and multiple channels for obtaining them (e.g., interviews, field observation, and iPad-based online search), and can collaborate with others to identify and collect user needs. **1**: With teacher support, students accomplish part of the objectives. **0**: With teacher support, students do not accomplish the objectives.

Stage	Scoring Criteria
Define	**4**: Students collaborate with others to investigate through multiple channels and design feasible solutions that overcome constraints, and provide concrete examples to demonstrate credibility. **3**: Students represent constraints using charts and tables based on expected goals. **2**: Students describe core concepts such as constraints and expected goals, and collaborate with others to categorize constraints into materials, time, cost, and individual capability. **1**: With teacher support, students accomplish part of the objectives. **0**: With teacher support, students do not accomplish the objectives.
Ideate	**4**: Students evaluate and synthesize others' viewpoints to design a concrete solution, and explain the reasons. **3**: Students compare and analyze solutions from the user and constraint perspectives, and list pros and cons. **2**: During in-group brainstorming, students propose different solutions and explain the rationale behind their ideas. **1**: With teacher support, students accomplish part of the objectives. **0**: With teacher support, students do not accomplish the objectives.
Prototype	**4**: After completing their assigned roles, students coordinate with and help other members; when disagreements arise, they handle them peacefully. **3**: Students fulfill role responsibilities and, in authentic contexts, use controlled variable methods to document iterations, changes, and adjustments to the design, experiment, or prototype. **2**: Students describe steps and methods for prototyping, collaborate via role division to build prototypes, and document iterations, changes, and adjustments to the design, experiment, or model. **1**: With teacher support, students accomplish part of the objectives. **0**: With teacher support, students do not accomplish the objectives.

Stage	Scoring Criteria
Share	**4**: Students proactively seek others' feedback, explore optimal improvement plans through multiple channels, and are willing to share ideas and products; they can identify others' strengths and weaknesses using feedback and provide reasonable suggestions. **3**: Students optimize the product after analyzing and evaluating feedback from users and others, and use data to justify optimization intentions; they can identify others' strengths and weaknesses and propose improvements. **2**: Students explain the core concept of optimization, and can propose improvement suggestions for others' work. **1**: With teacher support, students accomplish part of the objectives. **0**: With teacher support, students do not accomplish the objectives.
Total Score (Product Outcome)	**3**: The product addresses user needs and meets aesthetic and creativity standards. **2**: The product addresses user needs and meets aesthetic standards. **1**: The product addresses user needs. **0**: The product does not address user needs.

Table 14: Teacher rating rubric.

6.2.2 Metadata Indexing

To connect questionnaires, audio-video records, and teacher evaluations across sources, this study constructs a dual indexing mechanism that combines identity alignment and time alignment. Identity alignment uses anonymized student IDs (sid), project IDs (pid), and group IDs (gid) as unified keys. Time alignment uses event timestamps (t_event) and phase anchors as coordinates for locating process evidence. The indexing layer transforms multimodal raw evidence into a searchable evidence ledger, supporting subsequent sampling for annotation, feature extraction, and trace-back from model outputs. On the video side, the index records camera position, start and end times, phase anchors, and traceable pointers. On the process side, the index records milestone completion points and critical events (conflict, help-seeking, teacher intervention, and key iterations) to enable rapid retrieval of high-information segments. On the evaluation side, the index records scoring versions and stage standards, ensuring that cross-project comparisons do not suffer from implicit drift caused by minor rubric adjustments. With this mechanism, any structured field (e.g., an increase in conflict frequency, or a decrease in prototype-stage scores) can be traced back to the corresponding video segments and classroom context, forming a closed loop of model inputs, evidence basis, and explanation chain.

6.2.3 Data Annotation

Data annotation aims to convert theoretical constructs (e.g., collaboration quality, self-regulation, and iterative problem solving) into a variable system that is observable, computable, and verifiable. Considering resource constraints and executability in primary school classrooms, this study adopts a strategy of event-level lightweight annotation as the primary approach, complemented by denser annotation for selected key segments, and establishes an explicit mapping to teacher evaluation stages. Event-level annotation prioritizes process events that are strongly related to grouping: conflict, help-seeking, intervention, and iteration. For each event type, minimum decision conditions are specified. These annotations support process features such as conflict rate, distributions of help-seeking strategies, teacher support intensity, and iteration density, and are aligned with stage-level rubric standards such as prototyping and sharing, ensuring educational interpretability and consistency. Artifact and log annotation extracts structured fields from test records and engineering logs, complementing signals about debugging and reflection quality that are difficult to judge reliably from video alone. To ensure annotation reliability, the study recommends a quality control procedure based on double coding of sampled segments and agreement statistics, so that annotation outputs can serve as model inputs rather than only qualitative descriptions.

6.2.4 Standardized Packaging

After raw evidence preservation, metadata indexing, and multi-level annotation, all collected data are consolidated into a standardized dataset that is directly usable for intelligent grouping modeling and controlled evaluation. The packaged dataset follows a three-layer interface: *samples (features), labels, and metadata (evidence pointers)*. It is further aligned with graph-based formulations commonly adopted in grouping and learning analytics, where students are represented as nodes, peer relations as edges, and grouping outcomes (or teacher-evaluated performance) serve as supervision signals. Specifically, the sample layer includes node features that integrate questionnaire-derived learner priors (six maker-mindset dimensions) and individual process indicators to form student profiles, and group-level process features aggregated from event annotations (e.g., conflict, help-seeking, intervention, and iteration) to characterize collaboration dynamics and workflow efficiency. The label layer provides stage-wise and outcome-oriented rubric scores as evaluation targets for comparing control (P1 and P2) and treatment (P3 and P4) grouping conditions. Finally, the metadata layer retains time-stamped evidence pointers that link structured variables back to the original classroom context, enabling auditable trace-back and teacher-in-the-loop review. Table 15 below summarizes the resulting dataset schema and illustrates anonymized examples.

Data Layer	**Granularity**	**Example Fields**

Node Features (Student profiles)	Student	sid, mean thinking skills, mean self-efficacy, mean learning motivation, mean self-concept, mean self-regulation, mean collaborative learning.
		Real example (anonymized):
		`S01: 3.375, 3.600, 4.000, 3.600, 3.400, 4.000`
		`S02: 2.625, 4.800, 3.500, 2.400, 3.200, 3.167`

Group-Level Process Features (Derived from video event annotation)	Group and Project	Conflict rate, help-seeking frequency, teacher intervention count and type distribution, iteration count, and milestone efficiency (e.g., time to complete the prototype, time to start testing, and related indicators).
		Real example (anonymized):
		`P1, S03, S11, peer_rating=2, iteration count=87, conflict_cooccur=12, help_cooccur=20, teacher intervention count=35`
		`P2, S01, S05, peer_rating=4, iteration count=133, conflict_cooccur=0, help_cooccur=2, teacher intervention count=15`

Group-Level Process Features (Derived from audio event annotation)	Group and Project	pid, sid, t_event, content Real example (anonymized): `P4, teacher, s01, question="Why is the elevator not enough?", answer="Because there are too many dining cars."`

Supervision Labels (Derived from teacher rubric and observation sheet)	Group and Project	Stage-wise ratings: STEM knowledge and skills, empathize, define, ideate, prototype, and share (scores 4–0), and the overall product score (3–0).
		Real example (anonymized):
		`P3, G2, 3, 4, 3, 3, 4, 3, product=3`
		`P3, G4, 2, 3, 1, 3, 2, 2, product=1`

Metadata Pointers (For evidence trace-back)	Event and Phase	pid, gid, t_event, phase anchors (ideation, prototyping, testing, and sharing), event type (conflict, help-seeking, intervention, and iteration), and a brief evidence note, enabling trace-back from features and labels to the corresponding video context.
		Real example (anonymized):
		`P2, G1, event=Conflict, S02, S05, t_start=00:35:10, t_end=00:36:20`
		`P3, G2, event=help-seeking, S12, t_start=0:49:10, t_end=00:51:20`

Table 15: Standardized dataset schema.

7

Methodology for Intelligent Group Formation in PBL

The proposed methodology aims to generate fair, effective, and interpretable collaborative groups for primary school STEAM PBL under realistic classroom constraints. Rather than treating group formation as a generic clustering task, the framework formulates it as an evidence-driven decision process that explicitly accounts for multimodal uncertainty, interaction structure, and fixed group-size feasibility.

At a high level, the methodology consists of four sequential stages. First, heterogeneous classroom observations are abstracted into traceable evidence units at appropriate granularities, enabling consistent aggregation and auditability. Second, cross-modal alignment statistics are computed to quantify the reliability of transcript-based process evidence with respect to aligned audio-video context, and these statistics are propagated as reliability signals during feature construction. Third, a structure-aware interaction graph is constructed, and graph-based relational learning is applied to model peer influence and collaboration topology under reliability-aware semantics. Finally, a constrained multi-objective optimization module produces fixed-size grouping assignments that balance within-group complementarity, between-group equity, and collaboration potential, while generating evidence-based rationales that support teacher interpretation and intervention.

7.1 Problem Formulation

Let the student set be $\mathcal{S} = \{s_i\}_{i=1}^{N}$ with $N = 20$. For each project session $p \in \{1, 2, 3, 4\}$, let Δ_p denote the time window covering the full classroom activity (approximately 160 minutes). The goal is to output a grouping assignment $\pi^{(p)} : \mathcal{S} \to \{1, \ldots, K\}$ with $K = 5$ such that each group has exactly four students:

$$\forall g \in \{1, \ldots, K\}, \quad \left|\{s_i \in \mathcal{S} : \pi^{(p)}(s_i) = g\}\right| = 4. \tag{7.1}$$

We denote the set of feasible assignments by Ω, which minimally enforces the fixed group-size constraint and may additionally encode instructor-specified operational rules, such as must-link and cannot-link relations.

Group formation is formulated as a constrained multi-objective optimization problem:

$$\max_{\pi^{(p)} \in \Omega} \mathbf{F}(\pi^{(p)}) = \left[F_{\text{comp}}(\pi^{(p)}),\, F_{\text{collab}}(\pi^{(p)}),\, F_{\text{fair}}(\pi^{(p)}),\, F_{\text{stab}}(\pi^{(p)})\right]. \tag{7.2}$$

Here, F_{comp} measures within-group complementarity based on learner evidence, including relatively stable

trait indicators from questionnaires and process-level behavioral indicators derived from classroom discourse. F_{collab} captures interaction-dependent collaboration potential implied by observed peer communication structure during Δ_p and is evaluated through structure-aware relational modeling. F_{fair} quantifies between-group balance to avoid systematic disparities in the distribution of learner attributes and collaboration potential across groups. When multiple intelligent grouping sessions are conducted (Projects 3 and 4), F_{stab} penalizes unnecessary reassignment to promote temporal stability unless changes are justified by evidence.

A distinctive requirement in this study is that key quantities used by F_{comp} and F_{collab} are estimated from heterogeneous classroom evidence that is noisy, partially observed, and not uniformly reliable across students and time. Accordingly, subsequent sections introduce cross-modal alignment statistics that quantify the reliability of utterance-level evidence and provide signals for weighting, filtering, and propagating evidence contributions to downstream representations and interaction estimates. The output of the formulation, therefore, comprises not only the grouping assignment $\pi^{(p)}$ but also an auditable evidence rationale that traces objective-relevant signals back to traceable evidence units, enabling teacher interpretation and intervention.

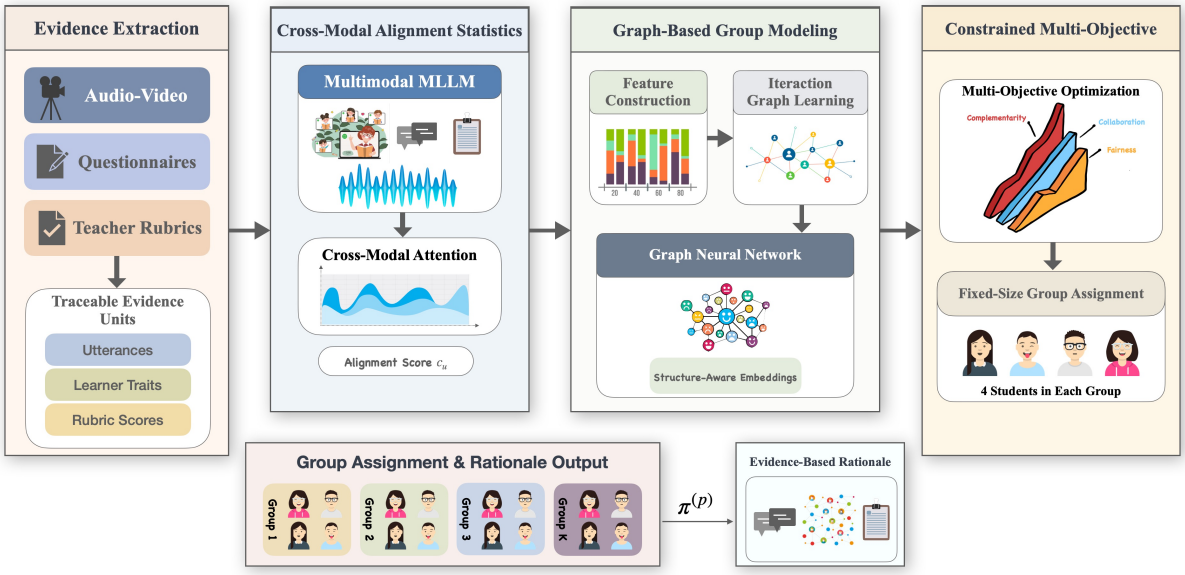


Figure 11: Overview of the proposed alignment-aware intelligent group formation framework.

As illustrated in Fig. 11, the proposed framework formulates group formation for elementary STEAM PBL as an alignment-aware, evidence-driven decision process rather than a purely data-driven clustering task. The framework first abstracts heterogeneous classroom data, such as audio–video recordings with aligned transcripts, learner questionnaires, and teacher rubrics, into traceable evidence units. Cross-modal alignment statistics are then computed using a frozen multimodal large language model to quantify the reliability of utterance-level transcript evidence with respect to its audio–video context, yielding alignment confidence scores ($c_u \in [0, 1]$) that are propagated as reliability weights in downstream aggregation.

On this basis, a directed, weighted interaction graph is constructed to model peer communication and collaboration structure, and a graph attention network is applied to learn structure-aware learner representations under reliability-aware edge semantics. Finally, a constrained multi-objective optimization module generates fixed-size group assignments that balance within-group complementarity, interaction-dependent collaboration potential, and between-group fairness, while producing evidence-based rationales that trace grouping decisions back to salient utterances, graph structures, and learner traits, thereby supporting interpretability and teacher-in-the-loop intervention.

7.2 Data Sources and Evidence Units

The proposed intelligent grouping framework is grounded in multiple sources of classroom data that capture complementary aspects of student learning and collaboration. Rather than treating these sources as raw inputs to a model, this study adopts an evidence-oriented perspective in which heterogeneous data are systematically abstracted into well-defined evidence units. Evidence units are treated as the minimal traceable carriers of information that can be aggregated, modeled relationally, and audited pedagogically, so that grouping outcomes remain interpretable and verifiable against concrete observations and instructional criteria. Three primary data sources are considered in each STEAM PBL project: classroom audio-video recordings with aligned transcripts, questionnaire-based learner self-reports, and teacher evaluation rubrics. Audio-video recordings document situated interaction processes as they unfold in real time, questionnaires capture relatively stable learner traits and dispositions, and teacher rubrics provide pedagogically grounded judgments of group performance and learning outcomes. Together, these sources form a multi-perspective evidence base that reflects both process and outcome dimensions of collaborative learning in authentic classroom settings, enabling the framework to support optimization while maintaining transparency, auditability, and teacher-in-the-loop refinement.

7.2.1 Classroom Audio-Video and Transcript Evidence

Classroom audio-video recordings are used to capture fine-grained interaction processes during project activities. To make these recordings computationally and analytically tractable, speech segments are transcribed and temporally aligned with the corresponding audio-video streams. Temporal alignment preserves the correspondence between linguistic content and classroom context, allowing process indicators to be grounded in observable events rather than decontextualized text. In this study, the transcript is treated as the primary carrier of process-level evidence because it provides a symbolic and indexable representation of classroom discourse while remaining linked to the original recordings via timestamps.

The basic unit of transcript-based evidence is defined as an utterance u, corresponding to a continuous speech segment produced by a single student within a bounded time interval. Each utterance is represented as $u = (p, uid, i, g, t_{\text{start}}, t_{\text{end}}, \text{text})$, where p denotes the project identifier, uid is a unique utterance identifier, i is the speaker identity, g denotes the local group context, and $(t_{\text{start}}, t_{\text{end}})$ are temporal boundaries. Utterance duration is defined as $\text{dur}(u) = t_{\text{end}} - t_{\text{start}}$. The utterance set of student i

in project p is denoted by $\mathcal{U}_i^{(p)}$, and these utterance-level units support aggregation of learner-level process indicators such as total speaking time $T_i^{(p)} = \sum_{u \in \mathcal{U}_i^{(p)}} \mathrm{dur}(u)$ and turn frequency $N_i^{(p)} = |\mathcal{U}_i^{(p)}|$. In addition to individual aggregation, utterance-level indexing by $(p, uid, t_{\mathrm{start}}, t_{\mathrm{end}})$ preserves traceability to the original recordings, enabling evidence-based rationales that can be audited against concrete classroom moments.

Because classroom recordings often contain noise, overlapping speech, and ambiguous alignment between transcripts and multimodal segments, this study incorporates a multimodal large language model as a frozen evidence extraction and alignment module. The multimodal model is not used as an end-to-end grouping policy; instead, it supports the construction of reliable and interpretable evidence signals from raw multimodal recordings. First, the model supports discourse-function labeling by mapping each utterance to an interpretable act label $\mathrm{act}(u) \in \mathcal{A}$ under a predefined taxonomy that captures pedagogically meaningful functions such as questioning, explaining, coordinating, providing feedback, documenting progress, presenting, and conflict repair. These labels instantiate transcript-derived behavioral evidence while remaining directly traceable to the underlying utterances.

Second, the multimodal model provides cross-modal alignment statistics that are summarized into an utterance-level confidence score $c_u \in [0, 1]$, representing the reliability of the alignment between the transcript segment and its corresponding audio–video evidence. Let A_u denote a normalized cross-modal attention distribution produced by the frozen model for utterance u. One generic construction defines the entropy $H(A_u) = -\sum_m A_u(m) \log(A_u(m) + \epsilon)$ and maps it to confidence via a monotonic transformation such as $c_u = \sigma(\alpha - \beta H(A_u))$, where $\sigma(\cdot)$ is a sigmoid and α, β are calibration constants. Methodologically, c_u is used as a reliability weight to downweight weakly aligned or noisy utterances during subsequent feature aggregation and interaction counting, improving robustness without introducing additional trainable parameters or risking overfitting in small-sample classroom settings. This positioning also preserves interpretability because the model outputs serve as auditable intermediate evidence signals rather than opaque decision rules.

7.2.2 Questionnaire-Based Learner Trait Evidence

To complement process-level observations, the study incorporates questionnaire-based self-reports collected before project implementation. The Maker Mindset pre-study questionnaire captures learner characteristics that are relatively stable over the short duration of the projects, including thinking skills, self-efficacy, learning motivation, self-concept, self-regulation, and collaborative learning orientation. These measures provide access to latent dispositions and self-perceived capabilities that are not directly observable from classroom discourse alone, and they supply a stable learner prior for complementarity and balance considerations in group formation.

Questionnaire responses are abstracted into learner-level evidence units by aggregating item scores within each conceptual dimension. For learner i and dimension d with item set I_d, the dimension score is

defined as $SUR_d(i) = \frac{1}{|I_d|}\sum_{j\in I_d} x_{i,j}^{\text{scored}}$, where $x_{i,j}^{\text{scored}}$ denotes the scored response after reverse-key correction when applicable. Reverse-keyed items are re-scored by $x_{i,j}^{\text{scored}} = (L+1) - x_{i,j}$ under an L-point Likert scale. The resulting trait evidence unit for learner i is the vector $\mathbf{u}_{i,\text{survey}} = [SUR_{\text{TS}}, SUR_{\text{SE}}, SUR_{\text{LM}}, SUR_{\text{SC}}, SUR_{\text{SR}}, SUR_{\text{CL}}]$, which is treated as stable across projects unless repeated measurements are available.

7.2.3 Teacher Evaluation Rubric Evidence

Teacher evaluation rubrics constitute a third source of evidence, providing an external pedagogical perspective on group performance and learning quality. The rubric used in this study assesses group work across multiple stages of the STEAM project process, including STEM knowledge and skills, empathize, define, ideate, prototype, share, and overall product outcome. These stage-based evaluations reflect instructional goals and criteria that are difficult to infer solely from behavioral data or self-reports, and they provide an outcome-oriented anchor for evaluating whether grouping decisions align with pedagogical intent.

Rubric scores are treated as group-level evidence units associated with each project. For project p and group g, the rubric evidence unit can be represented as $\mathbf{r}_g^{(p)} = [r_{\text{know}}, r_{\text{emp}}, r_{\text{def}}, r_{\text{ide}}, r_{\text{pro}}, r_{\text{sha}}, r_{\text{out}}]$, where the first six components are stage scores on a 0–4 scale and the outcome component is scored on a 0–3 scale. When a normalized overall rubric index is required for comparative evaluation, we define $R_g^{(p)} = \frac{1}{7}\left(\frac{r_{\text{know}}}{4} + \frac{r_{\text{emp}}}{4} + \frac{r_{\text{def}}}{4} + \frac{r_{\text{ide}}}{4} + \frac{r_{\text{pro}}}{4} + \frac{r_{\text{sha}}}{4} + \frac{r_{\text{out}}}{3}\right)$, which maps rubric evidence to $[0,1]$ while preserving stage contributions. Methodologically, rubric evidence is used for evaluation, interpretation, and teacher-facing auditing rather than as a direct optimization target during group formation, thereby avoiding circularity while still ensuring alignment with educational goals.

7.3 Cross-Modal Alignment Statistics

Cross-modal alignment statistics constitute a core methodological contribution of this study. Given the inherent noise and ambiguity in classroom recordings, alignment between transcript segments and their corresponding audio-video evidence are treated as a continuous and uncertain quantity rather than a binary condition. The objective of cross-modal alignment is to estimate the degree to which the semantic content of an utterance is supported by corresponding multimodal evidence.

For each utterance u, a frozen pre-trained multimodal large language model is used to compute cross-modal attention distributions between transcript tokens and aligned audio–video frames. Let A_u denote a normalized attention distribution aggregated across attention heads and layers. Alignment uncertainty is quantified by the entropy

$$H_u = -\sum_k A_u(k)\log(A_u(k)+\epsilon). \tag{7.3}$$

To obtain a comparable reliability measure, the alignment confidence score is defined as $c_u = \frac{H_{\max} - H_u}{H_{\max}}$, where $H_{\max}$ denotes the maximum attainable entropy under uniform attention. The score $c_u \in [0, 1]$ represents the reliability of utterance-level multimodal alignment and is used as a fundamental statistical signal throughout subsequent stages.

Across all data sources, a unifying principle of the proposed framework is the abstraction of raw data into evidence units that are both computationally manageable and pedagogically interpretable. Evidence units are defined at different granularities, including utterance-level units for transcript-based process evidence, learner-level units for questionnaire-based traits, and group-level units for rubric-based evaluations. Each evidence unit is explicitly linked to its source data through identifiers and temporal references, enabling traceability from high-level grouping decisions back to concrete classroom observations and assessment criteria. Concretely, utterance evidence is indexed by (p, uid) with timestamps and is augmented by multimodal-model-derived act labels $\mathrm{act}(u)$ and alignment confidence c_u. Questionnaire evidence is indexed by learner identity with dimension-level aggregation rules. Rubric evidence is indexed by project and group with stage-level score definitions. This traceability enables evidence-based rationales in which grouping decisions can be justified by referencing contributing evidence units rather than by opaque model outputs.

This evidence-centric design provides a coherent foundation for subsequent stages of the methodology. Feature engineering operates on evidence units rather than raw data, relational modeling builds on interaction relations derived from utterance-level evidence, and optimization balances objectives grounded in interpretable evidence constructs. As a result, the grouping framework not only supports effective and fair decision-making but also facilitates transparency, auditability, and pedagogical reflection in real classroom contexts.

7.4 Feature Engineering

This section operationalizes the evidence units defined by transforming questionnaire responses and transcript-based classroom observations into structured, decision-relevant indicators. The resulting learner representations provide the quantitative basis for modeling interaction-dependent collaboration potential and for optimizing group assignments under classroom constraints. In the subsequent section, these indicators are used to construct an interaction structure that explicitly captures peer communication patterns, enabling relational modeling and structure-aware grouping.

7.4.1 Learner-Level Behavioral and Trait Features

Learner-level trait features are derived from the Maker Mindset pre-study questionnaire and are treated as relatively stable within the short project horizon. For each learner i and each questionnaire dimension d, item responses are first transformed into scored values with consistent polarity and then aggregated into a dimension score. Let I_d denote the set of items belonging to dimension d, let $x_{i,j}$ denote the raw Likert

response of learner i on item j, and let I_{rev} denote the set of reverse-keyed items. The scored response is defined as $x_{i,j}^{\text{scored}} = (L+1) - x_{i,j}$ for $j \in I_{\text{rev}}$ and $x_{i,j}^{\text{scored}} = x_{i,j}$ otherwise, where L is the Likert scale length. The dimension score is computed as

$$SUR_d(i) = \frac{1}{|I_d|} \sum_{j \in I_d} x_{i,j}^{\text{scored}}. \tag{7.4}$$

This aggregation yields the trait vector $\mathbf{u}_{i,\text{survey}} = [SUR_{\text{TS}}, SUR_{\text{SE}}, SUR_{\text{LM}}, SUR_{\text{SC}}, SUR_{\text{SR}}, SUR_{\text{CL}}]$, corresponding to Thinking Skills, Self-Efficacy, Learning Motivation, Self-Concept, Self-Regulation, and Collaborative Learning. When missing responses occur, aggregation is performed under a controlled missingness policy suitable for small cohorts: *if the missing fraction within a dimension does not exceed a predefined threshold, the score is computed from available items; otherwise the dimension is treated as missing and imputed at the cohort level for modeling while retaining a missingness flag for analysis.* These trait features support complementarity and balance objectives by providing a stable characterization of individual differences that is not confounded by transient session dynamics.

Transcript-derived behavioral features quantify observable engagement and the functional nature of learner contributions during project activities. Let $\mathcal{U}_i^{(p)}$ denote the set of utterances produced by student i during project p within the time window Δ_p, and let $\text{dur}(u) = t_{\text{end}}(u) - t_{\text{start}}(u)$ denote utterance duration. Participation-related indicators are defined as

$$\begin{aligned} \text{PRC_SPK_TIME}(i,p) &= \sum_{u \in \mathcal{U}_i^{(p)}} \text{dur}(u), \\ \text{PRC_TURN_CNT}(i,p) &= |\mathcal{U}_i^{(p)}|, \\ \text{PRC_TURN_AVG}(i,p) &= \frac{\text{PRC_SPK_TIME}(i,p)}{\max(1, \text{PRC_TURN_CNT}(i,p))}. \end{aligned} \tag{7.5}$$

To capture relative contribution under fixed group size constraints, a within-group speaking-time share is computed. Let $\mathcal{G}_g^{(p)} = \{k : \pi^{(p)}(s_k) = g\}$ be the set of students assigned to group g in project p under a given grouping scheme and let $g_i = \pi^{(p)}(s_i)$ denote the group index of student i. The speaking-time share is

$$PRC_SPK_RATIO(i,p) = \frac{PRC_SPK_TIME(i,p)}{\sum_{k \in \mathcal{G}_{g_i}^{(p)}} PRC_SPK_TIME(k,p) + \epsilon}. \tag{7.6}$$

These indicators support both fairness diagnostics and collaboration readiness estimation by distinguishing balanced turn-taking patterns from dominance or disengagement.

To represent qualitative contribution beyond intensity, each utterance is assigned a primary discourse function label from an interpretable taxonomy $\mathcal{A}$. The act count and act rate for student i in project p

are defined as

$$CNT_a(i,p) = \sum_{u \in \mathcal{U}_i^{(p)}} \mathbb{I}[\mathrm{act}(u) = a], \quad RATE_a(i,p) = \frac{CNT_a(i,p)}{\max(1, PRC_TURN_CNT(i,p))}. \tag{7.7}$$

These indicators capture cognitive engagement, coordination, and regulation behaviors, accountability-related documentation, and learning-oriented social exchange. To improve robustness under sparse utterances, act-rate features can be smoothed via additive priors in implementation, which regularize categorical counts without altering the conceptual definitions.

Because transcript-derived indicators are obtained from multimodal classroom recordings, utterance reliability may vary due to noise, overlapping speech, or imperfect alignment across modalities. To mitigate the influence of unreliable evidence, confidence-aware aggregation is incorporated using cross-modal alignment statistics extracted from a frozen pre-trained multimodal model. For each utterance u, cross-modal attention distributions are summarized into a scalar alignment confidence score $c_u \in [0, 1]$, for example, by transforming the entropy of an attention distribution A_u via $H(A_u) = -\sum_m A_u(m) \log(A_u(m) + \epsilon)$ and a monotonic mapping such as $c_u = \sigma(\alpha - \beta H(A_u))$. Weighted aggregation replaces unweighted sums by confidence-weighted sums, yielding

$$PRC_SPK_TIME^w(i,p) = \sum_{u \in \mathcal{U}_i^{(p)}} c_u \, \mathrm{dur}(u), \quad CNT_a^w(i,p) = \sum_{u \in \mathcal{U}_i^{(p)}} c_u \, \mathbb{I}[\mathrm{act}(u) = a]. \tag{7.8}$$

The same weighting principle can be applied to response-related statistics so that low-confidence utterances contribute less to both learner indicators and interaction summaries. This mechanism is positioned as a reliability control and alignment cue rather than a direct explanation of grouping outcomes.

7.4.2 Interaction-Oriented Features, Normalization, and Structured Representation

Collaboration potential is interaction-dependent and therefore requires indicators that reflect responsiveness and mutual engagement. Using response relations inferred from conversational structure, responsiveness indicators are computed as

$$\begin{aligned} \mathrm{PRC_RESP_CNT}(i,p) &= \sum_{u \in \mathcal{U}_i^{(p)}} \mathbb{I}[\mathrm{is_response}(u) = 1], \\ \mathrm{PRC_RESP_RATE}(i,p) &= \frac{\mathrm{PRC_RESP_CNT}(i,p)}{\max(1, \mathrm{PRC_TURN_CNT}(i,p))}. \end{aligned} \tag{7.9}$$

In addition, pairwise interaction statistics summarize who responds to whom. For each ordered pair (i, j), the response count edge statistic is defined as

$$w_{ij}^{(p)} = \sum_{u:\, \mathrm{spk}(u)=i} \mathbb{I}[\mathrm{dst}(u) = j], \tag{7.10}$$

and an outgoing-normalized interaction tendency is defined as

$$\hat{w}_{ij}^{(p)} = \frac{w_{ij}^{(p)}}{\sum_{k \neq i} w_{ik}^{(p)} + \epsilon}. \tag{7.11}$$

At the learner level, node-wise summaries derived from $\{w_{ij}^{(p)}\}$ and $\{\hat{w}_{ij}^{(p)}\}$ may be included as interaction-oriented features, such as outgoing interaction mass, incoming response mass, and reciprocity proxies, providing a behavioral basis for distinguishing interactive participants from isolated speakers. At the relational level, the matrix $\{\hat{w}_{ij}^{(p)}\}$ serves as the empirical foundation for constructing an interaction structure in the subsequent section.

Because engineered features originate from different evidence sources and measurement scales, normalization is required to ensure comparability and stable downstream optimization. For each project p and each continuous feature dimension d, project-level standardization is defined as

$$z_{i,d}^{(p)} = \frac{x_{i,d}^{(p)} - \mu_d^{(p)}}{\sigma_d^{(p)} + \epsilon}, \tag{7.12}$$

where $\mu_d^{(p)}$ and $\sigma_d^{(p)}$ are computed across all learners for project p. Questionnaire traits may be standardized globally or within-project, depending on whether they are treated as stable priors or context-relative signals. In this study, standardization is applied consistently prior to downstream modeling to support multi-objective optimization. For interpretability, raw-scale values of bounded ratios may be retained for reporting and explanation, while standardized values are used for model-facing computations.

The final output of feature engineering is a structured learner representation for each project that integrates trait, process, and interaction-oriented indicators while preserving modular provenance for interpretability. The representation is defined as

$$\mathbf{u}_i^{(p)} = \left[\mathbf{u}_{i,\text{survey}}, \mathbf{u}_{i,\text{process}}^{(p)}, \mathbf{u}_{i,\text{interact}}^{(p)}\right], \tag{7.13}$$

where $\mathbf{u}_{i,\text{survey}}$ contains questionnaire-derived dimension scores, $\mathbf{u}_{i,\text{process}}^{(p)}$ contains transcript-derived participation and discourse-function indicators, and $\mathbf{u}_{i,\text{interact}}^{(p)}$ contains responsiveness and interaction-oriented summaries. This structured representation provides a direct interface to subsequent relational modeling and constrained optimization: *trait features support complementarity and balance objectives, process features capture engagement and collaboration readiness, and interaction summaries support interaction-dependent collaboration objectives and the construction of relational structures.* Because each component is computed from traceable evidence units, the representation also supports evidence-based rationales by enabling grouping decisions to be linked back to specific questionnaire dimensions and timestamped transcript segments.

7.5 Interaction Graph Construction

Group formation in collaborative STEAM activities is inherently relational because group effectiveness emerges from patterns of peer communication, responsiveness, and mutual influence rather than from individual attributes alone. Consistent with the interaction-dependent objective defined in Section 7.1 and the interaction-oriented indicators constructed in Section 7.4, we construct a directed weighted interaction graph for each project p to represent empirically observed peer interaction tendencies. The graph for project p is defined as $\mathcal{G}^{(p)} = (\mathcal{V}, \mathcal{E}^{(p)}, \mathbf{W}^{(p)})$, where $\mathcal{V} = \mathcal{S}$ is the student set, $\mathcal{E}^{(p)} \subseteq \mathcal{V} \times \mathcal{V}$ is the directed edge set, and $\mathbf{W}^{(p)}$ denotes the associated edge-weight matrix. A directed edge $(i, j) \in \mathcal{E}^{(p)}$ indicates that student i produced response utterances addressed to student j during project p, and the corresponding weight quantifies the strength of this interaction tendency within the project time window Δ_p. This representation provides an explicit interface between traceable transcript evidence and subsequent relational modeling, while remaining independent of any particular downstream graph learning architecture.

7.5.1 Edge Extraction from Reply Links

Edge extraction is grounded in utterance-level evidence units and their conversational linkages. Let $\mathcal{U}^{(p)}$ denote the set of all utterances observed in project p, and let $\mathrm{spk}(u)$ denote the speaker identity of utterance u. For response utterances, an addressee $\mathrm{dst}(u)$ is resolved using reply-to identifiers (e.g., `reply_to_uid`) and addressee tags when available; if explicit addressees are not present, the addressee is inferred by mapping `reply_to_uid` to the speaker of the referenced utterance. The raw directed interaction count from i to j in project p is defined as

$$w_{ij}^{(p)} = \sum_{u \in \mathcal{U}^{(p)}} \mathbb{I}[\mathrm{spk}(u) = i] \cdot \mathbb{I}[\mathrm{is_response}(u) = 1] \cdot \mathbb{I}[\mathrm{dst}(u) = j]. \tag{7.14}$$

This definition yields a traceable edge statistic in which each increment can be traced to a specific utterance identifier and timestamp, supporting auditability and evidence-based explanation.

Because learners differ in verbosity and turn frequency, raw interaction counts may overemphasize the contributions of highly talkative students. To mitigate this bias and obtain a normalized measure of interaction tendency, we compute an outgoing-normalized edge weight

$$\hat{w}_{ij}^{(p)} = \frac{w_{ij}^{(p)}}{\sum_{k \neq i} w_{ik}^{(p)} + \epsilon}, \tag{7.15}$$

where ϵ is a small constant for numerical stability. This normalization maps outgoing weights of each source node to a probability-like distribution over addressed peers, enabling comparison across students and supporting stable downstream modeling. When confidence-aware aggregation is enabled, each response utterance contribution can be weighted by its alignment confidence c_u so that low-confidence segments contribute less to $w_{ij}^{(p)}$, and its corresponding normalized form.

7.5.2 Multi-Evidence Edge Weight Fusion

While response-based interaction weights provide a direct measure of observed communication patterns, they may be complemented by additional evidence that captures latent affinity or pedagogically desirable complementarity. Accordingly, we define an optional multi-evidence fusion strategy in which the final edge weight is constructed as a convex combination of interaction tendency, trait similarity, and complementarity terms. Let $\mathbf{u}_{i,\text{survey}}$ denote the questionnaire-derived trait vector and let $\mathbf{u}_i^{(p)}$ denote the structured learner representation defined in Section 7.4. A fused edge weight is defined as

$$W_{ij}^{(p)} = \alpha\, \hat{w}_{ij}^{(p)} + \beta\, \mathrm{Sim}(\mathbf{u}_{i,\text{survey}}, \mathbf{u}_{j,\text{survey}}) + \gamma\, \mathrm{Comp}(\mathbf{u}_i^{(p)}, \mathbf{u}_j^{(p)}), \tag{7.16}$$

where $\alpha, \beta, \gamma \geq 0$ are hyperparameters controlling evidence contributions. The similarity term $\mathrm{Sim}(\cdot)$ may be instantiated as cosine similarity to reflect alignment of stable dispositions, while the complementarity term $\mathrm{Comp}(\cdot)$ is designed to reward heterogeneous yet compatible pairing patterns derived from trait and process indicators, such as balancing self-regulation and collaborative learning orientation across peers. This fusion mechanism is treated as optional because it introduces additional modeling assumptions beyond observed interaction. Therefore, in baseline Graph-Community experiments, we use $\hat{w}_{ij}^{(p)}$ only to ensure a clear ablation and to isolate the contribution of interaction structure derived from transcripts. In the proposed full framework, fusion can be enabled when the objective is to couple observed interaction tendencies with pedagogically motivated grouping preferences, while preserving interpretability through explicit term definitions and traceable evidence sources.

7.6 Graph Neural Modeling for Structure-Aware Grouping

This section introduces graph neural modeling as the relational inference component of the proposed grouping framework. Given the interaction graph constructed in Section 7.5 and the structured learner features constructed in Section 7.4, the purpose of graph neural modeling is to produce structure-aware representations that capture how observed communication patterns shape collaboration opportunities and peer influence. Unlike purely feature-based clustering, the graph-based representation explicitly conditions learner embeddings on neighborhood structure and edge weights, thereby aligning the modeling stage with the interaction-dependent collaboration objective defined in Section 7.1.

7.6.1 GNN Objective and Outputs

For each project p, let $\mathcal{G}^{(p)} = (\mathcal{V}, \mathcal{E}^{(p)}, \mathbf{W}^{(p)})$ denote the directed weighted interaction graph and let $\mathbf{u}_i^{(p)} \in \mathbb{R}^d$ denote the structured feature vector for student i. A graph neural network is used to compute a structure-aware embedding $\mathbf{g}_i^{(p)} \in \mathbb{R}^h$ for each student by iteratively aggregating information from neighbors under the interaction structure. At a high level, the GNN implements a message passing operator in which each node receives messages from its neighborhood, weighted by edge strength, and updates its representation through a learnable transformation. Formally, a generic L-layer message

passing formulation can be written as $\mathbf{h}_i^{(0)} = \mathbf{u}_i^{(p)}$ and, for $\ell = 0, \ldots, L-1$,

$$\mathbf{m}_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)} \phi^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, W_{ji}^{(p)}\right), \quad \mathbf{h}_i^{(\ell+1)} = \psi^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \mathbf{m}_i^{(\ell)}\right), \tag{7.17}$$

where $\mathcal{N}(i)$ denotes the neighborhood of i under the directed graph, $W_{ji}^{(p)}$ is the edge weight from j to i, and $\phi^{(\ell)}(\cdot)$ and $\psi^{(\ell)}(\cdot)$ are learnable functions that define message computation and node update, respectively. The final structure-aware embedding is defined as $\mathbf{g}_i^{(p)} = \mathbf{h}_i^{(L)}$.

The role of $\mathbf{g}_i^{(p)}$ in the grouping pipeline is twofold. First, structure-aware embeddings provide relational descriptors that can be used to generate candidate partitions or candidate neighborhoods for grouping, thereby improving stability and reducing search complexity under fixed group size constraints. Second, the embeddings support downstream evaluation of group compatibility by enabling computation of structure-informed similarity and cohesion measures that incorporate both learner attributes and interaction topology. In the full framework, these signals are used as inputs to the constrained optimization stage, complementing objective terms derived directly from feature statistics and interaction weights.

7.6.2 Attention-Based Aggregation and Structural Descriptors

To improve interpretability and to better capture heterogeneous peer influence patterns, we adopt attention-based aggregation as a preferred instantiation of the message passing operator. Under an attention-based GNN layer, each node i computes attention coefficients over its neighbors, and the aggregated message becomes a weighted combination of neighbor representations. Let $\alpha_{ij}^{(\ell,p)}$ denote the attention coefficient assigned by node i to neighbor j at layer ℓ in project p. A standard attention-based aggregation can be expressed as

$$\alpha_{ij}^{(\ell,p)} = \frac{\exp\left(e_{ij}^{(\ell,p)}\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(e_{ik}^{(\ell,p)}\right)}, \quad e_{ij}^{(\ell,p)} = \mathrm{Attn}^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, W_{ij}^{(p)}\right), \quad \mathbf{m}_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell,p)} \, \mathbf{W}^{(\ell)} \mathbf{h}_j^{(\ell)}, \tag{7.18}$$

where $\mathrm{Attn}^{(\ell)}(\cdot)$ is a learnable scoring function and $\mathbf{W}^{(\ell)}$ is a learnable linear map. Attention coefficients provide an explicit, quantitative description of how a learner's representation is influenced by different peers under the observed interaction structure, which supports subsequent evidence-based analysis at the relational level.

To obtain robust structure-aware descriptors suitable for grouping, we summarize attention distributions into compact statistics for each node. For a given node i, we define an attention-statistics vector $\mathbf{a}_i^{(p)}$ by aggregating the neighbor attention coefficients across heads and layers and then computing distributional summaries,

$$\mathbf{a}_i^{(p)} = \mathrm{Stat}\left(\{\alpha_{ij}^{(p)}\}_{j \in \mathcal{N}(i)}\right), \tag{7.19}$$

where $\alpha_{ij}^{(p)}$ denotes the aggregated attention assigned by i to j after averaging or pooling across attention heads and layers. The operator $\mathrm{Stat}(\cdot)$ includes concentration measures such as top-k attention mass, uncertainty measures such as attention entropy, and low-order moments that characterize dispersion. These statistics capture whether a learner's influence is broadly distributed across peers or concentrated on a small subset of collaborators, which is a meaningful structural characteristic in collaborative learning settings.

For candidate partition generation and stability enhancement, the attention-statistics vectors and/or the structure-aware embeddings are standardized within each project, optionally reduced by principal component analysis to retain a high fraction of variance, and then clustered to form candidate group prototypes. Specifically, we apply project-level Z-score normalization to the descriptor space, perform PCA to retain 90% cumulative explained variance, and apply K-means to obtain $K = 5$ clusters that serve as candidate grouping templates or initialization states. Subsequent constrained optimization refines these candidates to satisfy fixed-size constraints and to optimize the multi-objective criteria defined in Section 7.1.

7.7 Constrained Multi-Objective Group Optimization

This section specifies the decision-time optimization module that converts reliability-aware learner representations and interaction estimates into deployable group assignments under fixed classroom constraints. The optimization layer is intentionally separated from representation learning: preceding modules produce reliability-aware learner vectors and a reliability-aware interaction graph, while this module enforces classroom feasibility and balances multiple pedagogical objectives in an explicit, auditable manner.

7.7.1 Decision Variables and Feasible Set

Grouping decisions for project session p are represented by binary assignment variables. Let $x_{ig}^{(p)} \in \{0, 1\}$ indicate whether student s_i is assigned to group $g \in \{1, \ldots, K\}$, and let $\mathbf{X}^{(p)} = [x_{ig}^{(p)}]$ denote the assignment matrix, which is equivalent to the mapping $\pi^{(p)}$. The feasible set Ω is defined by hard classroom constraints. Each student must be assigned to exactly one group:

$$\sum_{g=1}^{K} x_{ig}^{(p)} = 1, \quad \forall s_i \in \mathcal{S}. \tag{7.20}$$

Each group must contain exactly four students:

$$\sum_{i=1}^{N} x_{ig}^{(p)} = 4, \quad \forall g \in \{1, \ldots, K\}. \tag{7.21}$$

Instructor-specified operational rules can be incorporated as additional hard constraints. For example, must-link constraints enforce co-assignment for designated student pairs, and can-not-link constraints prevent specific pairs from being co-assigned. These constraints are treated as feasibility conditions

rather than optimization preferences to ensure that all candidate solutions remain directly deployable in authentic classroom practice.

7.7.2 Objective Functions

The optimization seeks group assignments that balance multiple pedagogical objectives. Let $\mathbf{z}_i^{(p)}$ denote the reliability-aware learner representation produced by prior stages, and let $W_{ij}^{(p)}$ denote the directed reliability-aware interaction weight from student i to student j estimated from the session interaction graph. Four objectives are instantiated as scalar functions over $\mathbf{X}^{(p)}$, within-group complementarity, interaction-dependent collaboration potential, between-group fairness, and temporal stability.

Within-group complementarity. Complementarity is defined as functional diversity in the reliability-aware representation space and is encouraged by aggregating within-group pairwise complementarity scores:

$$F_{\text{comp}}(\mathbf{X}^{(p)}) = \sum_{g=1}^{K} \sum_{i<j} x_{ig}^{(p)} x_{jg}^{(p)} \, \text{Comp}(\mathbf{z}_i^{(p)}, \mathbf{z}_j^{(p)}). \tag{7.22}$$

The function $\text{Comp}(\cdot,\cdot)$ is designed to reward diverse yet compatible profiles rather than mere similarity, and it can be instantiated using distance-based, task-sensitive, or pedagogy-driven formulations.

Interaction-dependent collaboration potential. Collaboration potential is captured by accumulating reliability-aware within-group interaction strength:

$$F_{\text{collab}}(\mathbf{X}^{(p)}) = \sum_{g=1}^{K} \sum_{i \neq j} x_{ig}^{(p)} x_{jg}^{(p)} \, W_{ij}^{(p)}. \tag{7.23}$$

Because $W_{ij}^{(p)}$ is modulated by cross-modal reliability signals, interactions supported by weak or misaligned evidence contribute less to the objective, reducing sensitivity to classroom noise.

Between-group fairness. Between-group balance is enforced by penalizing deviations of aggregated group representations from the cohort-level mean. Let $A_g^{(p)} = \sum_i x_{ig}^{(p)} \mathbf{z}_i^{(p)}$ and $\bar{A}^{(p)} = \frac{1}{K} \sum_g A_g^{(p)}$. The fairness objective is:

$$F_{\text{fair}}(\mathbf{X}^{(p)}) = -\sum_{g=1}^{K} \left\| A_g^{(p)} - \bar{A}^{(p)} \right\|_2^2. \tag{7.24}$$

This term discourages systematic concentration of strengths or weaknesses within any single group and supports equitable distributions of learner attributes and collaboration potential.

Temporal stability. When intelligent grouping is performed across multiple sessions (Projects 3 and 4), stability is encouraged by penalizing unnecessary reassignment relative to the previous grouping $\pi^{(p-1)}$:

$$F_{\text{stab}}(\mathbf{X}^{(p)}) = -\sum_{i=1}^{N} \mathbb{I}\left[\pi^{(p)}(s_i) \neq \pi^{(p-1)}(s_i)\right]. \tag{7.25}$$

This objective promotes continuity without preventing evidence-justified changes when classroom conditions or inferred collaboration structures evolve.

7.7.3 Multi-Objective Aggregation and Solver

The multi-objective problem is converted into a scalar utility through weighted aggregation:

$$\max_{\mathbf{X}^{(p)} \in \Omega} \lambda_1 F_{\text{comp}}(\mathbf{X}^{(p)}) + \lambda_2 F_{\text{collab}}(\mathbf{X}^{(p)}) + \lambda_3 F_{\text{fair}}(\mathbf{X}^{(p)}) + \lambda_4 F_{\text{stab}}(\mathbf{X}^{(p)}). \tag{7.26}$$

The weights $\lambda_1, \ldots, \lambda_4$ encode instructional priorities and are fixed across sessions to ensure consistent decision policies, with sensitivity analysis reported in the experimental chapter. Because the cohort size is small ($N = 20$, $K = 5$), the optimization is computationally tractable and can be solved to optimality or near-optimality using integer programming solvers or efficient heuristic search, while preserving strict feasibility under Ω.

7.7.4 Auditability and Teacher Intervention

This optimization formulation supports interpretability by construction: each objective is computed explicitly, and the contribution of each term to the final decision can be inspected independently. Grouping rationales can be traced to learner representations $\mathbf{z}_i^{(p)}$, interaction evidence encoded in $W_{ij}^{(p)}$, and the reliability signals derived from cross-modal alignment statistics that modulate both representations and interaction weights. This design enables teachers to audit trade-offs among complementarity, collaboration potential, and fairness, and to intervene by adding or modifying constraints (e.g., must-link or cannot-link rules) or adjusting objective weights when classroom contingencies arise, thereby supporting a teacher-in-the-loop workflow rather than an opaque automated assignment.

8

Evaluation

This chapter evaluates the proposed classroom group formation framework that integrates reliability-aware multimodal representation learning, interaction-aware graph modeling, and constrained multi-objective optimization. The evaluation is conducted in a realistic classroom, where we assess both the effectiveness and equity of the resulting group assignments using normalized teacher rubric ratings that span the full design-thinking workflow (Knowledge and Skills, Empathize, Define, Ideate, Prototype, Share) as well as Product Outcome. To ensure that improvements are not merely descriptive but also mechanistically grounded, we additionally audit objective attainment using metrics that directly correspond to the optimization formulation (within-group complementarity, interaction-dependent collaboration potential, between-group fairness, and temporal stability). Together, these analyses establish an evidence chain from outcome-level gains to component-level contributions and decision-level auditability.

Section 8.1 presents the experimental setup, including dataset description and descriptive statistics, baseline and ablation definitions, and implementation details that support reproducibility. Section 8.2 reports the main results and provides a structured analysis, overall comparisons against practical baselines, stage-wise performance breakdowns, an analysis of the impact of cross-multimodal attention statistics, graph-based interaction modeling, and multi-objective optimization. Finally, Section 8.3 summarizes the key findings and implications for deployable, teacher-auditable grouping support in elementary STEAM PBL.

8.1 Experiments

8.1.1 Datasets

The evaluation is conducted on an authentic primary-school STEAM PBL classroom dataset collected across four projects (P1–P4) under a unified design thinking instructional structure. The analysis cohort comprises N = 20 elementary students, and grouping is performed with a fixed feasibility constraint of K = 5 groups of size 4 throughout the evaluation, matching the constrained assignment setting formalized in Chapter 7. Each project session is recorded as a continuous classroom timeline. Specifically, 160-minute audio–video recordings are collected per project, yielding 640 minutes of process evidence in total. The evaluation uses three complementary evidence sources: classroom audio–video recordings with temporally aligned transcripts, a pre-study learner questionnaire capturing relatively stable dispositions, and teacher rubric ratings providing pedagogically grounded outcome judgments.

Following the evidence-oriented data organization, heterogeneous raw observations are abstracted into traceable evidence units at different granularities. In addition to continuous project recordings, other modalities are shown as follows:

- **Utterance-level process evidence** Classroom discourse is represented as utterances indexed by project and timestamps. This unit is the minimal carrier for interaction modeling, reliability estimation, and teacher-facing trace-back, while it remains linked to the original audio–video streams through time anchors.
- **Learner-level trait evidence** A Maker Mindset pre-study questionnaire provides learner priors across six dimensions (thinking skills, self-efficacy, learning motivation, self-concept, self-regulation, and collaborative learning orientation). In the data collection, 20 valid questionnaire responses are archived with item-level granularity. In evaluation, we focus on the 20 students for whom the multimodal evidence chain is complete and consistently alignable across sources.
- **Group-level outcome evidence** Teacher evaluation rubrics rate group performance along related stages. The rubric includes 7 dimensions: six process/competency stages scored on a 0–4 scale (STEM knowledge and skills, empathize, define, ideate, prototype, share) and one product outcome dimension scored on a 0–3 scale. Rubric evidence is used for evaluation and auditing rather than being directly optimized as a training label, thereby reducing circularity while maintaining pedagogical validity.

Because classroom multimodal data are inherently noisy (overlapping speech, imperfect transcription, and occasional misalignment), evaluation is designed to remain robust under partial missingness and heterogeneous evidence quality. Reliability signals derived from cross-modal alignment statistics are used to downweight weakly supported utterances during aggregation and interaction counting, reducing the risk that incidental transcription or alignment errors dominate grouping decisions. When questionnaire dimensions have small missing fractions, cohort-level imputation is applied with explicit missingness flags retained for analysis, and sensitivity analyses are reported to assess whether conclusions are stable to these preprocessing choices.

To characterize the cohort and the evidence quality, Table 16 below summarizes the key dataset statistics used for evaluation reporting.

8.1.2 Baselines and Compared Methods

We compare the proposed framework with practical baselines commonly used in classrooms and with controlled ablations that isolate the contribution of each major component. All methods obey identical hard constraints, and all outcomes are evaluated using the same teacher rubric and the same set of objective attainment metrics. This ensures that improvements reflect method differences rather than feasibility or information advantages.

Statistic	Value
Cohort size	$N = 20$ students
Grouping setting	$K = 5$ groups, fixed size $= 4$
Number of projects	4 projects (P1-P4)
Audio-video recording (class view)	160 minutes per project
Audio-video recording (group view)	160 minutes per group per project
Total audio-video duration	3840 minutes
Questionnaire	40 valid responses
Questionnaire dimensions	6 (TS, SE, LM, SC, SR, CL)
Teacher rubric dimensions	7 (6 process + 1 outcome)

Table 16: Summary statistics of the classroom STEAM PBL evidence-chain dataset.

We include the following baselines, which cover both authentic practice and standard heuristic strategies:

- **Random grouping** Students are assigned uniformly at random with the same hard classroom constraints (fixed group size). This procedure is repeated for a large number of trials to construct a constraint-matched reference distribution, which is used for randomization-based statistical testing.

- **Teacher heuristic grouping** A teacher first proposes an initial grouping based on classroom experience (e.g., balancing perceived ability, behavior, and interpersonal compatibility).

- **Proposed intelligent grouping** This is the core grouping framework proposed in this work. It integrates reliability-aware multimodal evidence representation, interaction-aware graph-based learner modeling, and constrained multi-objective optimization. The method constructs learner representations from aligned multimodal evidence, models collaboration as a relational process via an interaction graph, and outputs a feasible grouping by jointly optimizing pedagogical objectives under the same classroom constraints.

- **Iterative intelligent grouping** This variant extends the proposed full framework by introducing an explicit cross-project update step. Specifically, learner representations and interaction weights are first learned using the evidence accumulated from P1 and P2. The system then incorporates P3 as an additional update episode, ingesting newly observed evidence (e.g., updated interaction traces, process logs, and teacher ratings when available) to refine the learner representations and re-estimate interaction strengths. After this update, the constrained multi-objective optimization is re-solved to produce the final grouping. This setting evaluates whether the framework can benefit from incremental evidence accumulation and whether grouping decisions remain stable and improve when an additional project episode is incorporated.

8.2 Results and Analysis

This section reports the main comparative evaluation across the four experimental conditions. All conditions operate under the same classroom constraint of fixed group size. The comparison is conducted in two complementary dimensions: effectiveness is measured by normalized teacher rubric ratings across design-thinking stages and Product Outcome, and equity is measured by between-group dispersion of overall rubric scores.

8.2.1 Overall Results

Table 17 summarizes the overall effectiveness and equity across the four grouping conditions. Random grouping yields an overall score of 0.63 ± 0.16 and exhibits the largest between-group disparity (variance 0.016, max–min gap 0.31), indicating that even under the same fixed-size constraint, feasible random partitions can produce highly uneven group outcomes and substantial uncertainty in expected performance. Teacher heuristic grouping improves the overall score to 0.70 ± 0.07 while markedly reducing dispersion (variance 0.003, gap 0.14), suggesting that teachers' constraint-aware experience is particularly effective in preventing extreme imbalances across groups. The proposed intelligent grouping achieves the highest mean overall score (0.80 ± 0.08), corresponding to an absolute gain of $+0.10$ over the teacher heuristic condition and $+0.17$ over random grouping. Its dispersion remains low (variance 0.004, gap 0.15), indicating that the framework improves mean outcomes without sacrificing group-level balance. Iterative intelligent grouping attains a comparable overall score of 0.79 ± 0.08, and further reduces the max–min gap to 0.12 (variance 0.004). Thus, suggesting that incorporating additional project evidence can enhance equity by tightening the spread of group outcomes, even when the mean performance is similar to the single-episode proposed setting. Overall, the results demonstrate a clear progression from chance-level grouping to teacher-informed grouping and to evidence-driven intelligent grouping, with intelligent methods achieving substantially higher average quality while maintaining low dispersion across groups.

Condition	Overall score (mean±95% CI)	Variance	Max-min gap
Random grouping	0.63 ± 0.16	0.016	0.31
Teacher heuristic grouping	0.70 ± 0.07	0.003	0.14
Proposed intelligent grouping	0.80 ± 0.08	0.004	0.15
Iterative intelligent grouping	0.79 ± 0.08	0.004	0.12

Table 17: Main comparison of overall effectiveness and equity.

While the above results establish outcome-level differences in both effectiveness and equity, they do not reveal which stages of the design-thinking workflow contribute most to the gains, nor how the improvements relate to the framework's internal decision criteria. We therefore proceed to a framework-level analysis, which decomposes performance across rubric dimensions and connects observed outcomes to decision-time objectives and evidence utilization.

8.2.2 Framework-level Performance

While the last section establishes outcome-level differences in overall effectiveness and equity, this subsection provides a framework-level interpretation of why these differences arise. Specifically, we triangulate evidence from five complementary views:

1. Stage-wise rubric diagnosis that decomposes the overall score into design-thinking dimensions.
2. Cross-multimodal attention/alignment statistics that quantify evidence reliability under realistic classroom noise and modality imbalance.
3. Interaction-graph structure statistics that characterize the relational signals available for graph-based learner modeling.
4. Constrained multi-objective optimization, balancing complementary pedagogical criteria such as within-group complementarity, interaction-dependent collaboration potential, and between-group fairness.
5. Cross-project group switching patterns that reveal stability and the value of iterative evidence accumulation.

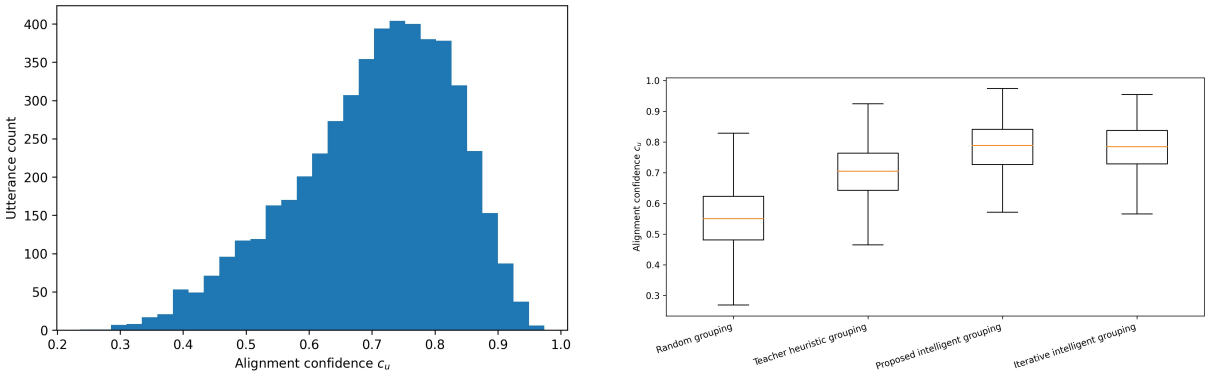
To diagnose which stages of the design-thinking workflow contribute to the overall differences, Table 18 reports dimension-wise normalized teacher rubric scores across STEM Knowledge and Skills, Empathize, Define, Ideate, Prototype, Share, and Product Outcome. Relative to Random grouping, the Teacher heuristic condition shows a broadly improved and more balanced profile, with notable gains in collaboration-execution stages (Prototype: 0.55→0.70; Share: 0.65→0.75), indicating reduced risk of weak coordination outcomes under instructor-led grouping.

Condition	STEM	Empathize	Define	Ideate	Prototype	Share	Product
Random grouping	0.60	0.65	0.65	0.65	0.55	0.65	0.67
Teacher heuristic grouping	0.65	0.70	0.70	0.70	0.70	0.75	0.67
Proposed intelligent grouping	0.80	0.80	0.75	0.80	0.80	0.75	0.87
Iterative intelligent grouping	0.75	0.85	0.80	0.80	0.75	0.75	0.80

Table 18: Dimension-wise normalized teacher rubric scores.

The proposed intelligent grouping attains consistently high scores across multiple stages (e.g., STEM = 0.80; Empathize/Ideate/Prototype = 0.80) and achieves the highest Product Outcome (0.87), substantially exceeding Random and Teacher heuristic (both 0.67). The Iterative intelligent grouping exhibits a different but competitive profile: Empathize increases to 0.85 and Define to 0.80, while Ideate and Share remain at 0.80 and 0.75, respectively, with a strong Product Outcome of 0.80. Overall, the dimension-wise decomposition supports the conclusion that the outcome-level differences are associated with coordinated

changes across several process dimensions and the final product score, rather than being driven by a single isolated rubric dimension.



Overall distribution of c_u c_u by grouping condition

Figure 12: Alignment confidence derived from cross-modal attention and alignment statistics.

A key design choice of the proposed framework is to operationalize cross-modal alignment and attention-derived statistics as an explicit reliability signal c_u, and to propagate this signal into both representation aggregation and interaction weighting. In authentic classrooms, transcripts can be noisy, modality availability is often imbalanced, and some utterances have weak cross-modal support, such as off-screen speech, overlapping talk, and occlusions. Figure 12 reports the overall distribution of the utterance-level alignment confidence c_u. The histogram is non-degenerate and skewed toward higher confidence values. Most utterances concentrate in the mid-to-high range, approximately from 0.60 to 0.85, while a low-confidence tail remains below 0.50. This pattern indicates that evidence reliability is heterogeneous at the utterance level rather than constant across the dataset, which motivates reliability-aware processing rather than uniform weighting. At the same time, the figure summarizes c_u across grouping conditions using boxplots. The median confidence increases from the Random grouping condition to the Teacher heuristic condition and remains high for the Proposed and Iterative intelligent grouping conditions. In addition, the interquartile ranges are tighter for the intelligent conditions, suggesting more consistently supported cross-modal evidence. Together, these results justify reliability-aware aggregation: the framework down-weights weakly supported utterances with low c_u and emphasizes segments with strong cross-modal support, thereby reducing sensitivity to incidental misalignment and partial missingness when constructing learner features and interaction edges for downstream graph modeling and constrained multi-objective optimization.

Beyond static learner attributes, the proposed framework models collaboration as a relational process through a student interaction graph constructed from observed dialogue links and participation traces. Rather than assuming a homogeneous interaction field, the within-group interaction patterns vary across projects and groups. Figure 13 below visualizes the intra-group networks for all four projects (P1–P4), where each subplot corresponds to one group within a given project. Across projects, two consistent structural characteristics emerge. First, the graphs exhibit pronounced within-group clustering: *students*

in the same team form relatively dense local neighborhoods, while interactions that connect across teams are rare by design, reflecting the small-group organization of STEAM PBL activities. Second, interaction evidence is heterogeneous both at the edge level and at the node level. Edge colors encode reliability-weighted, outgoing-normalized interaction intensity, showing that even within the same team, some student pairs are strongly supported by multimodal evidence while others are weak, intermittent, or absent. At the node level, several groups show asymmetric connectivity, where a subset of students concentrates a larger share of outgoing interaction mass through stronger or more numerous links, consistent with hub-like roles such as initiators, coordinators, or frequent responders.

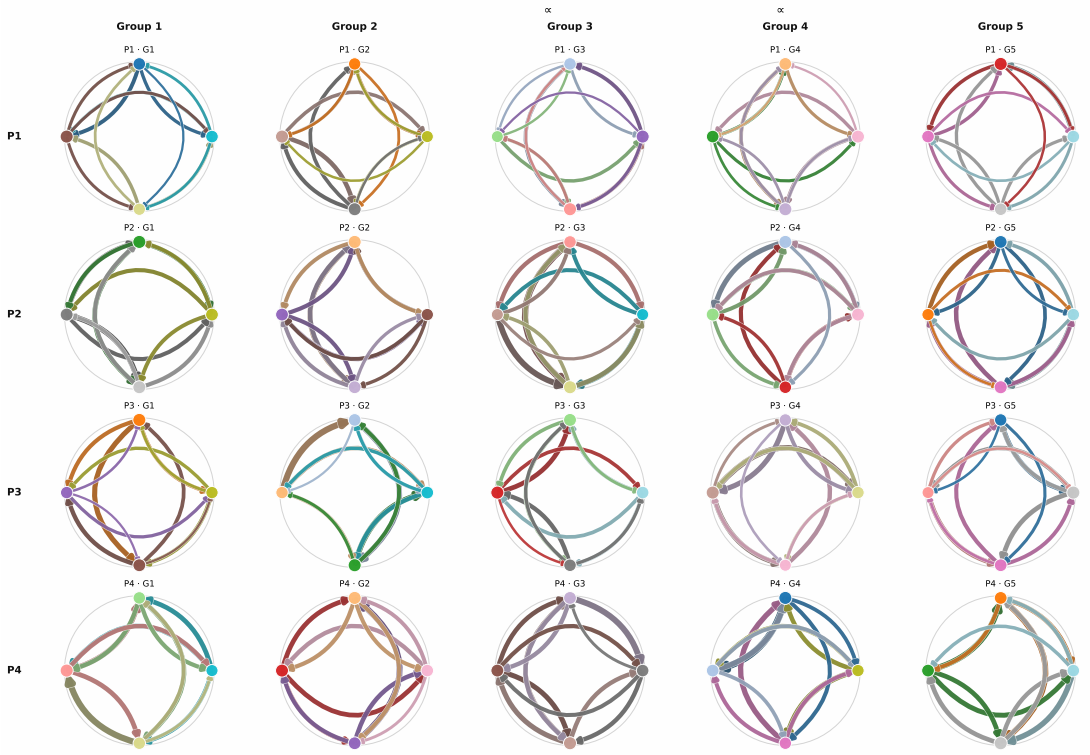


Figure 13: Group students interaction graph.

Importantly, the four-project visualization also highlights cross-project variability in how these roles and densities manifest. Some project episodes display more uniformly distributed within-group links, suggesting balanced participation, whereas others exhibit sharper centralization around one or two students, indicating coordination bottlenecks or dominance patterns. This combination of clustered topology and interaction heterogeneity provides direct motivation for graph neural networks: *message passing can exploit localized neighborhoods, edge-weighted connectivity, and role-dependent interaction structure to learn collaboration-aware learner representations that are not recoverable from independent profile features alone, and that remain robust when interaction evidence is sparse, or reliability is uneven.*

To complement aggregate grouping metrics, we evaluate framework-level performance in terms of its diagnostic fidelity for capturing heterogeneous learner growth across the six Maker-Mindset dimensions. The

four radar plots shown in Figure 14 serve as representative pre–post competency signatures, showing that the framework does not collapse development into a single gain score, but preserves multi-dimensional structure that teachers can interpret and trace back to evidence. More importantly, beyond broad-based and coherent improvement, the radar profiles surface risk-relevant divergences that directly motivate multi-objective grouping and teacher-in-the-loop adjustments. For example, S18 shows improved self-regulation and collaborative learning while learning motivation and self-concept decline, suggesting that behavioral adaptation in teamwork can co-occur with affective vulnerability and reduced academic self-perception. Such patterns justify framing grouping as a constrained decision problem that must balance performance, participation, and learner well-being. At the same time, this operationalizes framework-level performance as interpretability and instructional utility, rather than only predictive accuracy.



(a) Student S6,Group 1

(b) Student S10,Group 1

(c) Student S16,Group 1

(d) Student S18,Group 1

Figure 14: Pre-post radar profiles of students in proposed intelligent grouping.

While reliable multimodal representations and interaction-aware modeling provide informative signals,

the final grouping decision must be produced under hard classroom constraints and remain auditable to teachers. We therefore formulate grouping as a constrained multi-objective optimization problem, explicitly balancing within-group complementarity, interaction-dependent collaboration potential, and between-group fairness. To verify that the optimizer realizes these intended trade-offs in the finalized assignments, we compute objective-attainment metrics on the output groupings and report them in Table 19.

Condition	Complementarity	Collaboration potential	Fairness
Random grouping	0.20	0.65	0.65
Teacher heuristic grouping	0.21	0.72	0.83
Proposed intelligent grouping	0.23	0.78	0.87
Iterative intelligent grouping	0.21	0.80	0.85

Table 19: Objective-attainment auditing of the final groupings.

The auditing results align with the framework's design goals. Relative to Random grouping, all non-random conditions substantially improve collaboration potential and fairness (Random: 0.65/0.65). The Teacher heuristic baseline increases fairness to 0.83 and collaboration potential to 0.72, indicating that teachers can enforce equity-oriented constraints effectively but may not fully optimize for interaction-related outcomes. In contrast, the optimized conditions further raise collaboration potential (Proposed: 0.78; Iterative: 0.80) while also improving complementarity (Proposed: 0.23 versus 0.21 under Teacher heuristic). Notably, "Proposed intelligent grouping" achieves the best overall balance among the three objectives, attaining the highest complementarity (0.23) alongside strong fairness (0.87). "Iterative intelligent grouping" yields the highest collaboration potential (0.80) but with a small reduction in complementarity (0.21) and fairness (0.85) compared with "Proposed intelligent grouping", which is consistent with the expected trade-off surface when optimizing under the same hard constraints.

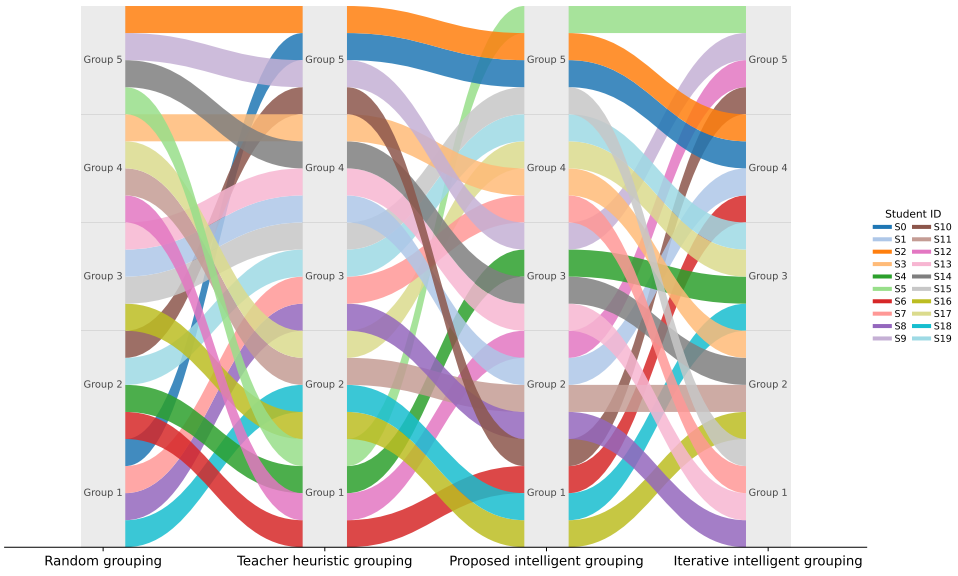


Figure 15: Group Scheme.

Group assignment is not fixed across projects in this dataset. To quantify the extent of cross-project reorganization, we derive a project-specific membership table by aggregating the utterance event log and assigning each student to the modal group label within each project. Figure 15 visualizes student-level group membership across projects, showing substantial reshuffling rather than minor local adjustments. Consistently, the student reassignment rate between consecutive projects is high, with 0.80 for P1$\rightarrow$P2, 0.75 for P2$\rightarrow$P3, and 0.85 for P3$\rightarrow$P4. This suggests that project episodes constitute meaningfully different collaboration contexts, motivating the explicit evaluation of temporal stability when iterative evidence accumulation is considered.

Beyond individual reassignment, we also audit stability at the level of peer relations and group composition. For peer retention, we compute the fraction of student pairs that are in the same group in the earlier project and remain co-members in the subsequent project. The pair retention rate decreases across transitions, from 0.1667 in P1$\rightarrow$P2 to 0.1333 in P2$\rightarrow$P3 and 0.0667 in P3$\rightarrow$P4, suggesting that repeated regrouping substantially alters peer exposure. For group-composition stability, we match groups across adjacent projects by maximum overlap and report the mean Jaccard similarity of matched group member sets. The mean Jaccard values are 0.3333 for P1$\rightarrow$P2, 0.2952 for P2$\rightarrow$P3, and 0.2190 for P3$\rightarrow$P4, indicating diminishing persistence of group structures over time.

Taken together, these results establish a strong empirical basis for incorporating stability-oriented analysis in evaluation. When the grouping framework is updated using additional project evidence, it should be assessed not only by mean performance and equity, but also by whether it avoids unnecessary reshuffling and preserves effective peer structures.

8.3 Summary

This chapter consolidates all experimental evidence for the proposed grouping framework, covering the dataset characteristics, evaluation protocol, and the full set of quantitative results. The evidence-chain dataset integrates synchronized classroom audio–video, utterance-level transcripts, process logs, pre–post questionnaires, and teacher rubric ratings. Empirically, the dataset exhibits non-trivial multimodal noise and missingness, and the cross-modal alignment confidence distribution is heterogeneous, with most utterances supported by mid-to-high reliability and a persistent low-confidence tail. This motivates the reliability-aware design of the framework, where alignment and attention-derived statistics are used as an explicit confidence signal to down-weight weakly supported segments during learner feature aggregation and to attenuate unreliable interaction evidence in edge construction.

Across the main comparison conditions, the results show consistent effectiveness and equity advantages for the intelligent grouping pipeline over constraint-matched random grouping and strong practice-based teacher grouping. In overall rubric performance, the proposed intelligent grouping attains the highest mean score while maintaining low disparity indicators. The dimension-wise rubric decomposition indicates that gains are distributed across multiple design-thinking stages rather than concentrated in a single rubric

dimension. The evaluation also provides decision-level evidence for constrained multi-objective optimization by auditing objective attainment on final groupings, demonstrating that improvements in teacher-rated outcomes coincide with stronger satisfaction of intended pedagogical criteria such as within-group complementarity, interaction-dependent collaboration potential, and between-group fairness under the same hard feasibility constraints. In addition, cross-project analyses reveal substantial regrouping dynamics: *student reassignment rates are high across consecutive projects, peer co-membership retention is low, and group-composition similarity decreases over time, establishing stability as a necessary evaluation axis for any iterative grouping strategy*. Complementary learner-state evidence from pre-post questionnaires shows overall positive shifts in multiple constructs, with the largest improvement observed in thinking-skills related measures, while collaboration-oriented self-report exhibits smaller change, supporting a triangulated interpretation of outcomes that is not solely dependent on a single measurement instrument.

Overall, the experimental dataset, outcome comparisons, mechanism-level diagnostics, and decision-level audits jointly support the thesis claim that reliable multimodal representation learning, interaction-aware graph modeling, and constrained multi-objective optimization form a coherent and accountable pipeline for classroom grouping. The evaluation demonstrates that the framework improves team outcomes in realistic settings while providing auditable evidence for why specific groupings are recommended and how performance-fairness-stability trade-offs can be controlled under explicit STEAM PBL classroom constraints.

9

Conclusions

This doctoral work investigates intelligent group formation for primary school STEAM PBL under realistic classroom conditions, where collaboration quality, equity, and feasibility must be balanced using heterogeneous and imperfect evidence. Departing from traditional heuristic-based or purely feature-driven grouping approaches, the thesis conceptualizes grouping as an evidence-to-decision process that integrates multimodal reliability modeling, interaction-aware learner representation, and constrained multi-objective optimization. The central premise is that effective and trustworthy grouping decisions cannot be derived solely from static student profiles or isolated performance indicators, but must instead account for the reliability of classroom evidence, the relational nature of collaboration, and the practical constraints governing instructional organization. By embedding these considerations into a unified and auditable framework, advances intelligent grouping from an ad hoc classroom practice toward a principled decision-support mechanism aligned with both learning analytics and pedagogical accountability.

This chapter concludes the study. Section 9.1 revisits the research questions and summarizes how the proposed framework answers them. Section 9.2 distills the primary scientific and practical contributions of the study. Section 9.3 discusses the main limitations and boundary conditions observed during development and evaluation, and Section 9.4 outlines promising directions for future research, including methodological extensions, longitudinal grouping policies, and responsible deployment considerations.

9.1 Responses to Research Questions

This study is structured around four research questions that collectively address how intelligent grouping can be realized in authentic educational settings:

RQ1 How can cross-modal alignment be quantified into alignment statistics that reflect the reliability of transcript evidence with respect to its aligned audio-video context, and how can such reliability signals be propagated to downstream learner representations under classroom noise and partial missingness?

RQ2 How can an interaction graph representation be constructed to capture students' collaborative relationships and complementary potential, while mitigating the influence of unreliable evidence and supporting structure-aware inference?

RQ3 How can an intelligent grouping mechanism be designed that integrates alignment-aware evidence signals with structure-aware graph modeling and produces fixed-size group assignments that balance within-group complementarity and between-group equity?

RQ4 How can objective functions and evaluation protocols be designed to ensure that grouping outcomes remain interpretable, traceable to evidence units, open to teacher intervention, and aligned with educational goals?

RQ1 asks how heterogeneous and noisy multimodal classroom evidence can be structured for grouping decisions. We address this by quantifying cross-modal alignment consistency as an explicit reliability signal and injecting it into both feature construction and relation/edge weighting, so that unreliable or weakly supported observations are down-weighted rather than propagated. The results indicate that reliability-aware modeling is a prerequisite for intelligent grouping in authentic classrooms, where multimodal streams differ markedly in quality, coverage, and semantic support. Treating alignment consistency as a first-class signal allows the grouping pipeline to privilege well-supported evidence, improving robustness to transcription noise, modality imbalance, and partial missingness.

RQ2 examines how to model collaborative learning potential beyond static profiles. Our solution represents the classroom as a student–student interaction graph, where edges encode directional responsiveness and peer influence, and node/edge statistics capture reciprocity and participation patterns. A structure-aware encoder (e.g., message passing over the interaction graph) then learns representations that summarize each student's relational position and collaborative role. Results indicate that these relational representations capture collaboration-relevant heterogeneity more effectively than attribute-only modeling, because responsiveness, reciprocity, and participation structure directly govern interaction opportunities and emergent group dynamics.

RQ3 asks how the learned student representations can be translated into classroom-valid groupings. We address this by formulating grouping as a constrained multi-objective optimization problem, where pedagogical goals are jointly optimized subject to hard feasibility constraints (e.g., fixed group sizes, assignment validity, and implementable grouping rules). The results indicate that pedagogically meaningful groupings arise from explicit trade-offs among multiple objectives, such as within-group complementarity, between-group fairness, and interaction balance, rather than from optimizing any single similarity or performance criterion. Moreover, the constrained multi-objective formulation provides a principled mechanism to encode classroom requirements for feasibility, balance, and stability directly in the search space, ensuring that optimized solutions remain actionable in real teaching settings.

RQ4 concerns the interpretability and practical adoption of algorithmic grouping. We address this by adopting interpretability-by-design: grouping decisions are generated through a transparent optimization structure, are traceable to underlying evidence and objective terms (with explicit contributions and reliability weights), and remain adjustable via teacher-defined constraints and preferences to support human-in-the-loop decision-making. The thesis further shows that interpretability in educational settings is most credible when embedded in the decision structure itself, so that each grouping can be explained in terms of evidence reliability, observed relational patterns, and objective-level trade-offs, enabling teachers to

inspect, revise, and justify group assignments within routine instructional practice.

9.2 Contributions

This study makes a set of concrete and interlocking contributions to intelligent classroom grouping and collaboration analytics, spanning evidence modeling, representation learning, decision optimization, and teacher-facing accountability.

At the level of multimodal evidence modeling, we propose a reliability-aware strategy for authentic classroom data by operationalizing cross-modal alignment consistency as an explicit indicator of evidence support strength. The resulting reliability signal is propagated into downstream feature construction and relational weighting, so that weakly supported observations are attenuated rather than amplified. This establishes a practical mechanism for improving grouping robustness under transcription noise, modality imbalance, and partial missingness, which are pervasive in real-world classroom capture and annotation pipelines.

At the level of learner representation, we advance an interaction-centric modeling approach that constructs student interaction graphs from classroom discourse and participation traces, and learns structure-aware representations that encode responsiveness, reciprocity, and influence. In this formulation, collaboration potential is treated as an emergent relational process rather than as a static function of individual profiles. Methodologically, the work consolidates graph-based relational inductive bias for learner modeling while maintaining stability under heterogeneous, sparse, and unevenly observed interactions.

At the level of group decision-making, we formulate classroom grouping as a constrained multi-objective optimization problem with explicit feasibility constraints (e.g., fixed group size and teacher-specified must-link/cannot-link rules) and separable pedagogical objectives (e.g., within-group complementarity and collaboration potential, between-group balance and equity, and cross-session stability). This formulation enables principled and auditable trade-offs among competing instructional goals, yielding grouping solutions that are both context-configurable and reproducible across implementations.

With respect to interpretability and practical adoption, the dissertation embeds accountability into the decision structure rather than treating explanation as a post-hoc add-on. Objective terms are decomposable, grouping rationales are traceable to evidence reliability and interaction structure, and teacher interventions can be incorporated through editable constraints or priority reweighting with predictable effects on feasibility and trade-offs. This supports a teacher-in-the-loop grouping paradigm in which algorithmic outputs are inspectable, adjustable, and justifiable within routine instructional practice.

In addition to these dissertation-specific contributions, the thesis is grounded in and extended by two methodological outcomes that reinforce the dissertation's core design commitments. The GEML framework provides a general mechanism for coupling language-based and graph-based learning streams through distribution-level mutual learning while keeping model architectures decoupled. By introducing

asymmetrical learning to address convergence-rate mismatch and uncertainty-aware coupling to stabilize cross-model knowledge transfer, GEML strengthens the dissertation's broader methodological stance that heterogeneous inductive biases should be integrated through explicit alignment mechanisms rather than implicit architectural entanglement. The CF-STGAT framework addresses non-IID heterogeneity in graph-based learning via attention-guided client clustering, cluster-wise optimization, and similarity-weighted inter-cluster fusion. CF-STGAT provides a principled template for stabilizing graph learning under structural and distributional divergence, aligning with the dissertation's emphasis on robustness and reliability when learning from heterogeneous relational data.

Collectively, these contributions articulate a coherent and transferable research program that integrates reliability-aware multimodal modeling, graph-structured representation learning, and constrained, interpretable decision optimization to support auditable and practically adoptable intelligent grouping in real classroom settings.

9.3 Limitations

Despite its contributions, this work has several limitations that should be considered when interpreting the results and assessing practical applicability. First, empirical validation is necessarily constrained by the realities of classroom research: the number of participating classes, schools, and projects is limited, and cohort sizes are bounded by class enrollment, which restricts statistical power and may reduce external validity across different grades, school cultures, curricular structures, and teacher practices. Relatedly, project-based learning outcomes are sensitive to contextual factors such as task design, teacher scaffolding intensity, classroom management style, and group history, making it difficult to fully disentangle grouping effects from instructional and contextual confounds, even under carefully controlled procedures.

Second, the proposed pipeline depends on the availability, completeness, and fidelity of multimodal evidence. While reliability-aware modeling mitigates noise, key upstream processes, including audio capture quality, speaker attribution, transcript accuracy, and the extraction of discourse or participation cues, can introduce systematic errors. In particular, multi-speaker overlap, child speech characteristics, and environmental noise may degrade transcript-derived interaction signals. Missing modalities or uneven coverage across students may lead to sparse or biased interaction graphs. Under such conditions, relational inference may become unstable, and learned representations may reflect observation artifacts (e.g., proximity to microphones or camera visibility) rather than true collaborative behaviors, which can, in turn, affect grouping recommendations.

Third, interaction graphs are abstractions of complex social processes and inevitably simplify classroom collaboration. The mapping from observed events (e.g., turn-taking, replies, mentions) to graph edges and weights may omit latent factors such as affective states, non-verbal coordination, off-task but socially supportive behavior, and evolving interpersonal relationships. Moreover, graph structure can be non-stationary across project phases, meaning that a representation learned in one stage may not transfer

perfectly to later stages without recalibration; this limitation is especially relevant in PBL settings, where roles, tasks, and participation norms shift over time.

Fourth, the constrained multi-objective formulation requires selecting objective definitions, constraint sets, and weightings that encode pedagogical priorities. Although this explicitness is a strength for transparency, it also introduces normative design choices: fairness can be operationalized as balance, compensation, or opportunity equity; stability can be framed as minimizing reassignment or supporting role rotation; and complementarity can emphasize cognitive diversity, skill diversity, or interaction compatibility. Different teachers and institutions may adopt different normative interpretations, and the study does not claim a universal objective specification; consequently, responsible use requires local calibration and stakeholder alignment. The sensitivity of outcomes to weight settings should be further characterized in broader deployments.

Finally, translating the framework into large-scale, routine classroom use would require additional engineering and governance beyond the study's scope. Multimodal classroom data is highly sensitive, particularly when minors are involved. Robust deployment would require end-to-end privacy and compliance measures, including informed consent procedures, data minimization, secure storage and access control, audit logging, retention policies, and clear institutional governance regarding acceptable use. In addition, adoption depends on usability constraints: teachers must be able to understand outputs, apply constraints efficiently, and trust that recommendations are stable and justifiable within limited preparation time. These human factors, together with infrastructure variability across schools, represent practical limitations that warrant dedicated future work.

9.4 Future Work

Future research can extend this study from complementary perspectives, spanning classroom experimentation, methodological advancement, and real-world deployment. From the perspective of classroom studies and empirical validation, future work should involve larger-scale and multi-site classroom experiments across different grade levels, school contexts, and instructional designs to assess robustness and generalizability. Longitudinal classroom studies that track students across multiple projects or academic terms would enable systematic investigation of how grouping strategies influence not only short-term collaboration quality, but also longer-term outcomes such as role development, participation trajectories, social inclusion, and learning equity. Such studies would also allow grouping to be examined as an adaptive process that evolves in response to prior group experiences, teacher interventions, and changing project demands, rather than as an isolated decision made at the start of each activity.

From a methodological perspective, several extensions are promising. More advanced uncertainty-aware and self-supervised multimodal alignment techniques could further strengthen reliability estimation under severe noise, partial observability, or modality loss, which are common in authentic classroom environments. Interaction modeling could be enriched by incorporating non-verbal behavioral cues, artifact-level

collaboration traces, and temporal role dynamics through multi-layer or time-evolving graphs, enabling learner representations that better reflect the full complexity of collaborative processes. In addition, learning adaptive objective weighting or preference models from historical teacher decisions and classroom outcomes could reduce the burden of manual calibration and improve alignment between optimization criteria and pedagogical intent.

From a decision and optimization perspective, future work could move beyond single-instance grouping toward longitudinal and policy-based grouping strategies. Such strategies would explicitly manage trade-offs among stability, diversity, and opportunity equity over extended periods, allowing constraints on repeated pairings, controlled role rotation, and cumulative exposure to collaborative opportunities. Integrating grouping decisions with downstream instructional interventions, such as role assignment, scaffolding prompts, or collaboration scripts, would further connect grouping with actionable pedagogical support rather than treating it as a standalone organizational task.

From a systems and deployment perspective, developing richer teacher-facing interfaces remains a critical direction. Interactive tools that support counterfactual exploration (e.g., visualizing how groupings change under different constraints or priorities), explanation of objective trade-offs, and incorporation of teacher feedback as preference signals would strengthen human–AI collaboration in instructional planning. Finally, responsible and scalable deployment will require privacy-preserving and resource-aware implementations, including edge-based feature extraction, federated or split learning approaches, and data-minimization strategies that respect institutional constraints while maintaining interpretability and decision quality.

Beyond the educational setting, an additional and important direction is to transfer the proposed core methodology to other collaborative project domains that face analogous grouping challenges. Many real-world contexts, such as corporate cross-functional project teams, software engineering teams, clinical multidisciplinary teams, research consortia, volunteer coordination, and online collaborative communities, require forming small groups under heterogeneous evidence, relational dynamics, and operational constraints. The study's pipeline of reliability-aware evidence modeling, interaction-graph-based representation learning, and constrained multi-objective grouping can serve as a domain-agnostic template that is instantiated with domain-specific evidence sources, interaction definitions, and objective constraints. Future work can therefore investigate cross-domain adaptation strategies, including how to redefine interaction graphs in non-classroom environments, how to calibrate reliability signals when modalities differ (e.g., logs, emails, task boards, or sensor traces), and how to design objectives that reflect domain goals (e.g., productivity, safety, fairness, resilience, or compliance). Such extensions would clarify the generality of the proposed framework and contribute to a broader theory and practice of accountable intelligent grouping in collaborative systems.

Bibliography

[1] K. Bf. “Framework for 21st century learning definitions”. In: *Battelle for Kids* (2019) (cit. on p. 2).

[2] R. M. Gillies. “Cooperative learning: Review of research and practice”. In: *Australian Journal of Teacher Education (Online)* 41.3 (2016), pp. 39–54 (cit. on pp. 2, 12).

[3] X. Ochoa et al. “Multimodal learning analytics-Rationale, process, examples, and direction”. In: *The handbook of learning analytics* 2 (2022), pp. 54–65 (cit. on p. 4).

[4] L. Chen et al. “Tackling uncertain correspondences for multi-modal entity alignment”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 119386–119410 (cit. on p. 4).

[5] B. Barron. “When smart groups fail”. In: *The journal of the learning sciences* 12.3 (2003), pp. 307–359 (cit. on p. 4).

[6] B. Flanagan et al. “Towards explainable group formation by knowledge map based genetic algorithm”. In: *2021 International conference on advanced learning technologies (ICALT)*. IEEE. 2021, pp. 370–372 (cit. on p. 4).

[7] S. Wu. “Application of artificial intelligence graph convolutional network in classroom grade evaluation”. In: *Scientific Reports* 15.1 (2025), p. 32044 (cit. on p. 4).

[8] I. Mateos-Aparicio-Ruiz et al. “Spatio-temporal graph neural networks for human-AI collaborative decision-making”. In: *Machine Learning with Applications* (2025), p. 100771 (cit. on p. 4).

[9] S. Garshasbi et al. “Optimal learning group formation: A multi-objective heuristic search strategy for enhancing inter-group homogeneity and intra-group heterogeneity”. In: *Expert systems with applications* 118 (2019), pp. 506–521 (cit. on p. 5).

[10] P. B. Miranda, R. F. Mello, and A. C. Nascimento. “A multi-objective optimization approach for the group formation problem”. In: *Expert Systems with Applications* 162 (2020), p. 113828 (cit. on pp. 5, 24, 28).

[11] L. Zhang et al. “Forming a robust team in educational scenarios using genetic algorithm with partial repair operators”. In: *Education and Information Technologies* (2025), pp. 1–26 (cit. on p. 5).

[12] E. Perignat and J. Katz-Buonincontro. “STEAM in practice and research: An integrative literature review”. In: *Thinking skills and creativity* 31 (2019), pp. 31–43 (cit. on p. 9).

[13] K. Chappell et al. “A framework for effective STEAM education: Pedagogy for responding to wicked problems”. In: *International Journal of Educational Research Open* 9 (2025), p. 100474 (cit. on p. 9).

[14] P. C. Blumenfeld et al. “Motivating project-based learning: Sustaining the doing, supporting the learning”. In: *Educational psychologist* 26.3-4 (1991), pp. 369–398 (cit. on p. 9).

[15] J. Krajcik and P. Blumenfeld. “19. Project-Based Learning”. In: *The Cambridge handbook of the learning sciences* (2006), pp. 317–333 (cit. on p. 9).

[16] J. W. Thomas. *A review of research on project-based learning*. 2000 (cit. on p. 9).

[17] C.-H. Chen and Y.-C. Yang. “Revisiting the effects of project-based learning on students’ academic achievement: A meta-analysis investigating moderators”. In: *Educational Research Review* 26 (2019), pp. 71–81 (cit. on p. 10).

[18] B. Condliffe. “Project-Based Learning: A Literature Review. Working Paper.” In: *MDRC* (2017) (cit. on p. 10).

[19] J. Krajcik et al. “Assessing the effect of project-based learning on science learning in elementary schools”. In: *American Educational Research Journal* 60.1 (2023), pp. 70–102 (cit. on p. 10).

[20] T. R. Kelley and J. G. Knowles. “A conceptual framework for integrated STEM education”. In: *International Journal of STEM education* 3.1 (2016), p. 11 (cit. on p. 10).

[21] A. T. Estapa and K. M. Tank. “Supporting integrated STEM in the elementary classroom: a professional development approach centered on an engineering design challenge”. In: *International Journal of STEM education* 4.1 (2017), p. 6 (cit. on p. 10).

[22] E. A. Dare, J. A. Ellis, and G. H. Roehrig. “Understanding science teachers’ implementations of integrated STEM curricular units through a phenomenological multiple case study”. In: *International journal of STEM education* 5.1 (2018), p. 4 (cit. on p. 10).

[23] N. R. Council et al. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press, 2012 (cit. on p. 10).

[24] M. Pedaste et al. “Phases of inquiry-based learning: Definitions and the inquiry cycle”. In: *Educational research review* 14 (2015), pp. 47–61 (cit. on p. 10).

[25] D. F. Bjorklund and A. D. Pellegrini. “Child development and evolutionary psychology”. In: *Child development* 71.6 (2000), pp. 1687–1708 (cit. on p. 11).

[26] J. Mandara. “The typological approach in child and family psychology: A review of theory, methods, and research”. In: *Clinical child and family psychology review* 6.2 (2003), pp. 129–146 (cit. on p. 11).

[27] A. K. Tolmie et al. “Social effects of collaborative learning in primary schools”. In: *Learning and instruction* 20.3 (2010), pp. 177–191 (cit. on pp. 11, 12).

[28] Z. Zhan et al. “Effects of gender grouping on students’ group performance, individual achievements and attitudes in computer-supported collaborative learning”. In: *Computers in Human Behavior* 48 (2015), pp. 587–596 (cit. on p. 11).

[29] M. Witt, B. Knight, and T. Booth. “Personality-sensitive pedagogies: A study of small group interactive behaviours among 9-to 10-year-olds”. In: *British Educational Research Journal* 50.6 (2024), pp. 2875–2898 (cit. on p. 11).

[30] P. Kutnick and A. Kington. “Children’s friendships and learning in school: Cognitive enhancement through social interaction?” In: *British journal of educational psychology* 75.4 (2005), pp. 521–538 (cit. on p. 11).

[31] S. Hallam, J. Ireson, and J. Davies. “Grouping practices in the primary school: what influences change?” In: *British Educational Research Journal* 30.1 (2004), pp. 117–140 (cit. on p. 12).

[32] A. M. van Dijk, T. H. Eysink, and T. de Jong. “Supporting cooperative dialogue in heterogeneous groups in elementary education”. In: *Small Group Research* 51.4 (2020), pp. 464–491 (cit. on p. 12).

[33] S. Kim et al. “Influence of teachers’ grouping strategies on children’s peer social experiences in early elementary classrooms”. In: *Frontiers in psychology* 11 (2020), p. 587170 (cit. on p. 12).

[34] M. M. Hendrickx et al. “Social dynamics in the classroom: Teacher support and conflict and the peer ecology”. In: *Teaching and teacher education* 53 (2016), pp. 30–40 (cit. on p. 12).

[35] J. L. Webb-Williams. “Teachers’ use of within-class ability groups in the primary classroom: A mixed methods study of social comparison”. In: *Frontiers in psychology* 12 (2021), p. 728104 (cit. on p. 12).

[36] V. Battistich, D. Solomon, and K. Delucchi. “Interaction processes and student outcomes in cooperative learning groups”. In: *The Elementary School Journal* 94.1 (1993), pp. 19–32 (cit. on p. 12).

[37] P. Blikstein and M. Worsley. “Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks”. In: *Journal of learning analytics* 3.2 (2016), pp. 220–238 (cit. on pp. 14, 25).

[38] K. Mangaroska et al. “Multimodal Learning Analytics to Inform Learning Design: Lessons Learned from Computing Education.” In: *Journal of Learning Analytics* 7.3 (2020), pp. 79–97 (cit. on pp. 14, 25).

[39] M. Worsley, R. Martinez-Maldonado, and C. D’Angelo. “A New Era in Multimodal Learning Analytics: Twelve Core Commitments to Ground and Grow MMLA.” In: *Journal of Learning Analytics* 8.3 (2021), pp. 10–27 (cit. on p. 14).

[40] P.-C. Sun et al. “A design to promote group learning in e-learning: Experiences from the field”. In: *Computers & Education* 50.3 (2008), pp. 661–677 (cit. on p. 20).

[41] C. Liang, R. Majumdar, and H. Ogata. “Learning log-based automatic group formation: system design and classroom implementation study”. In: *Research and Practice in Technology Enhanced Learning* 16.1 (2021), p. 14 (cit. on p. 20).

[42] J. Moreno, D. A. Ovalle, and R. M. Vicari. “A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics”. In: *Computers & Education* 58.1 (2012), pp. 560–569 (cit. on pp. 20, 24, 27).

[43] Z. Sun and M. Chiarandini. “An exact algorithm for group formation to promote collaborative learning”. In: *LAK21: 11th International Learning Analytics and Knowledge Conference*. 2021, pp. 546–552 (cit. on p. 20).

[44] C. Odo, J. Masthoff, and N. Beacham. “Group formation for collaborative learning”. In: *International conference on artificial intelligence in education*. Springer. 2019, pp. 206–212 (cit. on p. 20).

[45] N. Maqtary, A. Mohsen, and K. Bechkoum. “Group formation techniques in computer-supported collaborative learning: A systematic literature review”. In: *Technology, Knowledge and Learning* 24.2 (2019), pp. 169–190 (cit. on pp. 23, 24, 27).

[46] W. M. Cruz and S. Isotani. “Group formation algorithms in collaborative learning contexts: A systematic mapping of the literature”. In: *CYTED-RITOS International Workshop on Groupware*. Springer. 2014, pp. 199–214 (cit. on pp. 23, 24, 27, 28).

[47] T. R. Henry. “Creating effective student groups: an introduction to groupformation. org”. In: *Proceeding of the 44th ACM technical symposium on Computer science education*. 2013, pp. 645–650 (cit. on p. 23).

[48] C. E. Christodoulopoulos and K. A. Papanikolaou. “A group formation tool in an e-learning context”. In: *19th IEEE international conference on tools with artificial intelligence (ICTAI 2007)*. Vol. 2. IEEE. 2007, pp. 117–123 (cit. on p. 23).

[49] Y.-T. Lin, Y.-M. Huang, and S.-C. Cheng. “An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization”. In: *Computers & Education* 55.4 (2010), pp. 1483–1493 (cit. on pp. 24, 27).

[50] M. C. Rosatelli and J. A. Self. “A collaborative case study system for distance learning”. In: *International Journal of Artificial Intelligence in Education* 14.1 (2004), pp. 97–125 (cit. on p. 24).

[51] N. M. Hewahi and E. A. Hamra. “A hybrid approach based on genetic algorithm and particle swarm optimization to improve neural network classification”. In: *Journal of Information Technology Research (JITR)* 10.3 (2017), pp. 48–68 (cit. on pp. 24, 28).

[52] I. Srba and M. Bielikova. “Dynamic group formation as an approach to collaborative learning support”. In: *IEEE transactions on learning technologies* 8.2 (2014), pp. 173–186 (cit. on p. 24).

[53] M. Ullmann, D. Ferreira, and C. Camilo-Junior. “An automatic group formation method to promote student interaction in distance education courses”. In: *Research Anthology on Developing Critical Thinking Skills in Students*. IGI Global Scientific Publishing, 2021, pp. 1064–1085 (cit. on p. 24).

[54] C.-M. Chen and C.-H. Kuo. “An optimized group formation scheme to promote collaborative problem-based learning”. In: *Computers & Education* 133 (2019), pp. 94–115 (cit. on p. 24).

[55] H. Li et al. “Collaborative Learning Groupings Incorporating Deep Knowledge Tracing Optimization Strategies”. In: *Applied Sciences* 15.5 (2025), p. 2692 (cit. on pp. 24, 27, 28).

[56] P. Blikstein. “Multimodal learning analytics”. In: *Proceedings of the third international conference on learning analytics and knowledge*. 2013, pp. 102–106 (cit. on p. 25).

[57] D. Di Mitri et al. “From signals to knowledge: A conceptual model for multimodal learning analytics”. In: *Journal of Computer Assisted Learning* 34.4 (2018), pp. 338–349 (cit. on p. 25).

[58] H. Ouhaichi, D. Spikol, and B. Vogel. “Research trends in multimodal learning analytics: A systematic mapping study”. In: *Computers and Education: Artificial Intelligence* 4 (2023), p. 100136 (cit. on p. 25).

[59] S. Wang et al. “Artificial intelligence in education: A systematic literature review”. In: *Expert Systems with Applications* 252 (2024), p. 124167 (cit. on p. 25).

[60] B. Dong et al. “Large language models in education: A systematic review”. In: *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*. IEEE. 2024, pp. 131–134 (cit. on p. 25).

[61] S. Küchemann et al. “On opportunities and challenges of large multimodal foundation models in education”. In: *npj Science of Learning* 10.1 (2025), p. 11 (cit. on p. 25).

[62] J. Zhou et al. “Graph neural networks: A review of methods and applications”. In: *AI open* 1 (2020), pp. 57–81 (cit. on p. 26).

[63] B. Khemani et al. “A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions”. In: *Journal of Big Data* 11.1 (2024), p. 18 (cit. on p. 26).

[64] H. Nakagawa, Y. Iwasawa, and Y. Matsuo. “Graph-based knowledge tracing: modeling student proficiency using graph neural network”. In: *IEEE/WIC/aCM international conference on web intelligence*. 2019, pp. 156–163 (cit. on p. 26).

[65] M. Li et al. “Multi-view hypergraph neural networks for student academic performance prediction”. In: *Engineering Applications of Artificial Intelligence* 114 (2022), p. 105174 (cit. on p. 26).

[66] Q. Huang and Y. Zeng. “Improving academic performance predictions with dual graph neural networks”. In: *Complex & Intelligent Systems* 10.3 (2024), pp. 3557–3575 (cit. on p. 26).

[67] S. Hakkal and A. Ait Lahcen. “Leveraging graph neural network for learner performance prediction”. In: *Expert Systems with Applications* (2025), p. 128724 (cit. on p. 26).

[68] H. Ma et al. “Dgcd: An adaptive denoising gnn for group-level cognitive diagnosis”. In: *The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)*. 2024 (cit. on p. 27).

[69] H. Acosta et al. “Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning”. In: (2024) (cit. on p. 27).

[70] H. Acosta et al. “A Multi-View Predictive Student Modeling Framework with Interpretable Causal Graph Discovery for Collaborative Learning Analytics.” In: *International Educational Data Mining Society* (2025) (cit. on p. 27).

[71] L. Zheng et al. “A novel group cognitive graph approach for improving collaborative problem solving”. In: *International Journal of Educational Technology in Higher Education* 22.1 (2025), p. 18 (cit. on p. 27).

[72] D. Roh et al. “SIG-Net: GNN based dropout prediction in MOOCs using Student Interaction Graph”. In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. 2024, pp. 29–37 (cit. on p. 27).

[73] M. R. Ullmann et al. “Formation of learning groups in cmoocs using particle swarm optimization”. In: *2015 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2015, pp. 3296–3304 (cit. on p. 27).

[74] Y. Zheng et al. “An improved genetic approach for composing optimal collaborative learning groups”. In: *Knowledge-Based Systems* 139 (2018), pp. 214–225 (cit. on p. 27).

[75] Z. Zhang, Z. Zhou, and Y. Wang. “SSEGCN: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis”. In: *Proceedings of the 2022 conference of the North American Chapter of the association for computational linguistics: human language technologies*. 2022, pp. 4916–4925 (cit. on p. 30).

[76] M. Pan et al. “SPRF: A semantic Pseudo-relevance Feedback enhancement for information retrieval via ConceptNet”. In: *Knowledge-Based Systems* 274 (2023), p. 110602 (cit. on p. 30).

[77] S. Rizou et al. “Efficient intent classification and entity recognition for university administrative services employing deep learning models”. In: *Intelligent Systems with Applications* 19 (2023), p. 200247 (cit. on p. 30).

[78] S. I. Wang and C. D. Manning. “Baselines and bigrams: Simple, good sentiment and topic classification”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012, pp. 90–94 (cit. on p. 30).

[79] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 (cit. on p. 30).

[80] G. Forman. “BNS feature scaling: an improved representation over tf-idf for svm text classification”. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008, pp. 263–270 (cit. on p. 30).

[81] M. E. Maron. “Automatic indexing: an experimental inquiry”. In: *Journal of the ACM (JACM)* 8.3 (1961), pp. 404–417 (cit. on p. 30).

[82] T. Joachims. “Text categorization with support vector machines: Learning with many relevant features”. In: *European conference on machine learning*. Springer. 1998, pp. 137–142 (cit. on p. 30).

[83] Y. Kim. “Convolutional neural networks for sentence classification”. In: *arXiv preprint arXiv:1408.5882* (2014) (cit. on p. 30).

[84] T. Mikolov et al. “Recurrent neural network based language model.” In: *Interspeech*. Vol. 2. 3. Makuhari. 2010, pp. 1045–1048 (cit. on p. 30).

[85] J. Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186 (cit. on p. 30).

[86] Y. Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019) (cit. on p. 30).

[87] L. Yao, C. Mao, and Y. Luo. “Graph convolutional networks for text classification”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 7370–7377 (cit. on p. 31).

[88] Z. Lu, P. Du, and J.-Y. Nie. “VGCN-BERT: augmenting BERT with graph embedding for text classification”. In: *European conference on information retrieval*. Springer. 2020, pp. 369–382 (cit. on p. 31).

[89] G. Hinton, O. Vinyals, and J. Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015) (cit. on p. 31).

[90] L. Hu, S. Liu, and W. Feng. “Skeleton-based action recognition with local dynamic spatial–temporal aggregation”. In: *Expert Systems with Applications* 232 (2023), p. 120683 (cit. on p. 41).

[91] B. McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282 (cit. on p. 42).

[92] T. Li et al. “Federated optimization in heterogeneous networks”. In: *Proceedings of Machine learning and systems* 2 (2020), pp. 429–450 (cit. on p. 42).

[93] Q. Li, B. He, and D. Song. “Model-contrastive federated learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 10713–10722 (cit. on p. 42).

[94] S. P. Karimireddy et al. “Scaffold: Stochastic controlled averaging for federated learning”. In: *International conference on machine learning*. PMLR. 2020, pp. 5132–5143 (cit. on p. 42).

[95] J. Wang et al. “Tackling the objective inconsistency problem in heterogeneous federated optimization”. In: *Advances in neural information processing systems* 33 (2020), pp. 7611–7623 (cit. on p. 42).

[96] F. Sattler, K.-R. Müller, and W. Samek. “Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints”. In: *IEEE transactions on neural networks and learning systems* 32.8 (2020), pp. 3710–3722 (cit. on p. 43).

[97] R. Presotto, G. Civitarese, and C. Bettini. “Fedclar: Federated clustering for personalized sensor-based human activity recognition”. In: *2022 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE. 2022, pp. 227–236 (cit. on p. 43).